北京大學

本科生毕业论文

题目：基于机器学习的笛子演奏技法识别研究与实现

Research and implementation of Chinese flute playing technique recognition based on machine learning

姓　　名：马英浩

学　　号：1600010740

院　　系：数学科学学院

本科专业：数学与应用数学

指导导师：陈晓鸥

二〇二〇 年 五 月

北京大学本科毕业论文导师评阅表

<table>
<tr><td>学生姓名</td><td>马英浩</td><td>学生学号</td><td>1600010740</td><td>论文成绩</td><td></td></tr>
<tr><td>学院（系）</td><td colspan="3">数学科学学院</td><td>学生所在专业</td><td>数学与应用数学</td></tr>
<tr><td>导师姓名</td><td>陈晓鸥</td><td>导师单位/
所在研究所</td><td>北京大学王选计算机研究所</td><td>导师职称</td><td>研究员</td></tr>
<tr><td colspan="2">论文题目
（中、英文）</td><td colspan="4">基于机器学习的笛子演奏技法识别研究与实现
Research and implementation of Chinese flute playing technique recognition based on machine learning</td></tr>
<tr><td colspan="6">导师评语
（包含对论文的性质、难度、分量、综合训练等是否符合培养目标的目的等评价）
该生毕业设计题目是基于机器学习的笛子演奏技法识别研究与实现，设计内容是运用 FCNN 全卷积神经网技术实现对笛子演奏技法的自动识别，该题目难易程度和工作量适中，能体现综合训练的要求。毕业设计中使用自己构建的笛子演奏技法数据集，采用音频事件检测的策略，通过 Pytorch 深度学习工具，搭建了一个 FCNN 事件检测系统，训练了一个笛子演奏技法识别模型，实验结果显示了该模型的有效性。毕业论文条理比较清楚、文理基本通顺、书写格式规范，反映了该生专业基础知识比较扎实，初步具备了综合运用所学知识分析问题、解决问题的能力。

导师签名：

2020 年 5 月 27 日</td></tr>
</table>

版权声明

摘要

音乐对象识别和记录是音乐信息检索的基本内容。与旋律提取和音乐转录的其他领域不同，有关乐器技法检测的研究仍处于早期阶段，现有的工作主要集中于单个音符或逐帧的技法检测，并且缺乏有效的数据集。本文构建了关于竹笛 10 种技法的 101 分钟音频数据集，提取 mel 谱作为音频特征，并提出了一种基于全卷积神经网络（FCNN）的不定长输入音频的事件检测的端到端方法，用于乐器演奏技术的检测。本文将其与基于 VGG13、VGG16 和 LeNet-5 等基线模型比较，在数据集上测试评估该框架的有效性，得到了最高为 94%的平均准确率，并且验证集上全部技法均有较高准确率，输入不同长度的音频都有较高准确率，具有一定的泛化能力。

关键词：音频事件检测、音乐信息检索、全卷积神经网络、演奏技法鉴别

Abstract

Music object recognition and recording is the essential component of music information retrieval. Different from other fields of melody extraction and music transcription, the research on musical instrument technique detection is still in the early stage. The existing work mainly focuses on technique detection of individual notes or frame, and there is a lack of effective data set. This article constructed audio dataset on ten kinds of techniques of transverse Chinese bamboo flute (Di), with 101 minutes of audio, using Mel frequency spectrum as audio feature, and put forward a model based on the fully convolutional neural network (FCNN). This model is an end-to-end sound event detector for variable length of the input and can be used in the detection of instruments technology. Compared to the model based on VGG13, VGG16 and LeNet-5 baseline to evaluate the effectiveness of the proposed framework, this model got the average accuracy 94%, much better than all the others, on input of the different length of audio, which suggest good generalization ability.

Key Words： sound event detection, music information retrieval, fully convolutional neural network, musical instruments technique detection

全文目录

第一章 引言

1.1研究背景和意义

音频事件是指在一段音频信号中具有特定语义的，有时限的，对应于特定动作活动或过程的音频信号片段[1]，是人类可以通过人耳来识别和判定的声音事件。相比较其他音频事件或声学信号，例如汽车鸣笛声、敲门声、语音信号等，音乐声是一类比较特殊的音频：其可能有多个声音来源的各个成分在时间域和频率域上都有一定的相关性，不宜直接套用音频事件检测或自动语音识别的方法，这使得在音乐信息检索领域中，识别音乐场景、音乐对象，并将音乐音频信号转化为某种形式的音乐符号记录下来，成为一个很有挑战的任务。它包括几个子任务：基频估计、起止时间检测、乐器鉴别、节拍节奏追踪、技法检测等等。由于识别记录音乐信号所包含的子任务数量多、应用范围广，该任务在有挑战性的同时也被认为是音乐信息检索领域的一个基本问题[2][3]。

成熟的音乐信息识别记录系统将有很广泛的应用，包括音乐辅导教育、自动音乐伴奏、音乐制作（音乐可视化智能编辑），音乐搜索（按照旋律、节奏、技法、和声等标签）等等[4]。因此，音乐信息识别、记录（谱）有很大潜在的经济效益和社会影响。而这样的技术又和其他的音乐信息检索任务密切相关，例如声部分离、结构划分、乐曲检测、音乐相似性评估等等[5][6]。可以说，音乐信息识别和记谱衔接了音乐音频和音乐符号，对音乐的进一步理解、音乐版权维护和基于乐谱的自动作曲有很大的帮助。

对中国音乐而言，乐器类别、乐器基频和起止时间都和西方乐器相近，有一些关于钢琴或提琴的研究可以效仿，但中国音乐经常出现的自由节奏和复杂技法，则有待进一步的研究。人们在技法检测领域的研究远远不及转谱，但现有的转谱方式在特殊技法出无法有效运行。有许多有用的音乐转录方法，例如通过 F0 、CQT 等[7][8]，但它们主要关注独奏乐器的旋律提取，对于特殊技法，例如颤音，由于它一帧包含大量不同的音符，因此每个音符的时间值非常短。旋律提取算法中的后处理阈值或其他参数难以有效提取此时的频率信息，因此某些技法的特殊音高或其他特征可能会被忽略，因而会丢失一些重要信息，无法准确转录音乐。

这样的不准确或许在一些西方音乐中无关紧要，但事实上，中国音乐的“韵味”常常体现在一个骨干音到另一个骨干音之间的变化过程，或体现在乐句结尾漫长的拖腔中。这对鉴别音乐流派及版本鉴别，口传文化和有声文化的记录、保护和研究，民族音乐的学习与传承都至关重要的意义价值。

实现从音频中记录检测演奏技法还可以对音乐产业和教育应用产生变革性的影响。诸如基于演奏技术检测的音频内容分析和音乐辅助教学（包括乐器演奏或音乐作曲中的配乐课程）等应用程序可能是很好的例证。

相对于声乐、戏剧，乐器往往存在更严格的律制和更复杂的技法。古人有“丝不如竹，竹不如肉”的观点，认为竹笛这种气鸣类乐器通过人的嘴吹奏，演奏位置最接近人的心脏，是最能体现人的“心声”的乐器。而竹笛经过在中国七八千年的发展，诞生了纷繁复杂的技法，很多技法有很强的个人风格色彩或地域流派色彩。总地来说，北方的流派常用舌和指的技巧，例如花舌、飞指、颤音、顿音、历音，旋律分割，节奏感强；南方流派擅长气的技巧，例如气震音、指颤音、泛音、倚音、叠音、连音等，旋律连贯、平稳、柔和，具有抒情性[9]。很多特殊技法，例如仿马叫、掌揉音、压揉音等只被特定的流派甚至演奏家在部分音乐作品中使用。可以说，研究竹笛的技法对于研究中国音乐的技法乃至记谱有重要的意义。

1.2本文主要研究内容

本文使用深度学习的相关方法，从特征提取、模型训练与检测等方面，系统地研究了竹笛单旋律技法检测的问题。考虑到所选竹笛技法通常一次只出现一种，我们期望直接构建一个多分类器，提高竹笛技法检测的准确率。

与旋律提取和音乐转录的其他方面不同，关于演奏技术检测的研究仍处于早期阶段。这项任务的困难之一在于数据采集和数据标记。在所有乐器中，气鸣类乐器的演奏技法丰富，发声受很多因素影响，难以为表演者检测和手动标记，因此在该领域的研究数量也不多。为此，本文构建了关于竹笛技法的数据集。

在本文中，我们通过深度学习方法提出了一种基于全卷积神经网络（FCNN）[10]的单声道声音事件检测端到端方法，用于中国竹笛演奏技法检测。该模型通过短时傅里叶变换计算不同技法的 mel 谱用于训练。该模型可以一次性接受长度为

10 秒的音频，最终输出 10 秒钟每一秒的技法判别信息，并将结果和逐秒判别的基线模型比较，证明了 FCNN 模型相对优异的表现，填补了有关研究领域的空白，为进一步的多声部技法鉴别、音乐记谱等任务做了铺垫。

有关的 Python3 代码和完整的数据集（参见第三章）发布在 nicolaus625.github.io 上。

1.3本文组织结构

正文第一章为引言，简要阐述了音乐对象识别和记录的研究背景和意义，并阐述了乐器技法检测特别是竹笛技法检测对中国音乐识别和记录的意义。总结了本课题的研究方向和主要研究内容。

第二章对国内外相关领域的研究现状做了梳理总结，回顾了有关的音频特征提取和深度学习模型理论基础。最后介绍了可用的数据集。

第三章介绍了本文采用的关于 9 种中国竹笛常见的技法音频和无技法音频的带标注数据集的构建过程。

第四章提出了基于 FCNN 等深度学习方法的竹笛技法检测模型，介绍了利用 mel 特征，在 VGG 模型之后构建转置卷积层回复各频段和每一秒的预测，并直接给出输入每一秒技法判别的方法，比较了 FCNN 和其他模型的结果。

结尾对本文进行了总结，并展望了未来可继续研究的有关工作。

第二章 相关工作

乐器技法检测本质上是机器学习问题，该问题首先需要将音频数据降维，提取有鉴别能力的特征，然后建立并训练合适的分类器。从已知数据中提取有效特征后利用模型训练学习知识，再利用知识对未知数据进行预测。近年来深度神经网络整合了特征提取和分类器训练，但考虑到音频信号丰富的变化性和样本量少等问题，大部分神经网络模型目前仍需要通过预处理获得基本特征而后训练模型。预处理通常包括采样率归一化、分帧、加窗短时傅里叶变换等。

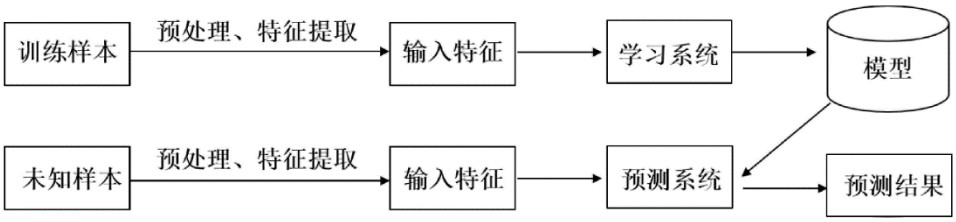

图1 音频信息检索机器学习模型基本框架

2.1关于音频特征提取的研究

音频信号是根声道数、采样率、时间长度等参数有关的一系列数据，特征提取就是从其分帧后的小片段数据中提取到有效向量表示。时域统计特征较为简单，例如短时平均能量、短时平均幅度、过零率等等，因而大部分音频信号的时域特征鉴别能力较弱，很少单独使用，在深度学习模型中很少作为模型输入[11]。

在频率域，有三种基于人听觉系统近似对数感知特性的特征常被用于音频事件检测：梅尔频带能量（mel）、对数梅尔频带能量（Logmel）和梅尔频率倒谱系数（MFCC）[12]。Xia 等人结合动态调整阈值的方法提取音频的 Mel 特征进行音频事件检测。Hayashi 等人采用 Logmel 特征作为模型的输入。Heittola 等人利用 MFCC 特征进行音频事件检测，Lu 等人则提取了包含 MFCC 在内的多个特征进行组合[1]。虽然传统特征特别是 MFCC 或 Logmel 在传统模型上有一定的表现，特别是 Logmel 被广泛应用于传统信号处理中重叠音频检测，但也有鲁棒性不足

等问题[12]，同时深度学习神经网络模型通常将高级特征提取内化为网络模型训练的一部分，通常只采用短时傅里叶变换后最原始的 mel 特征。

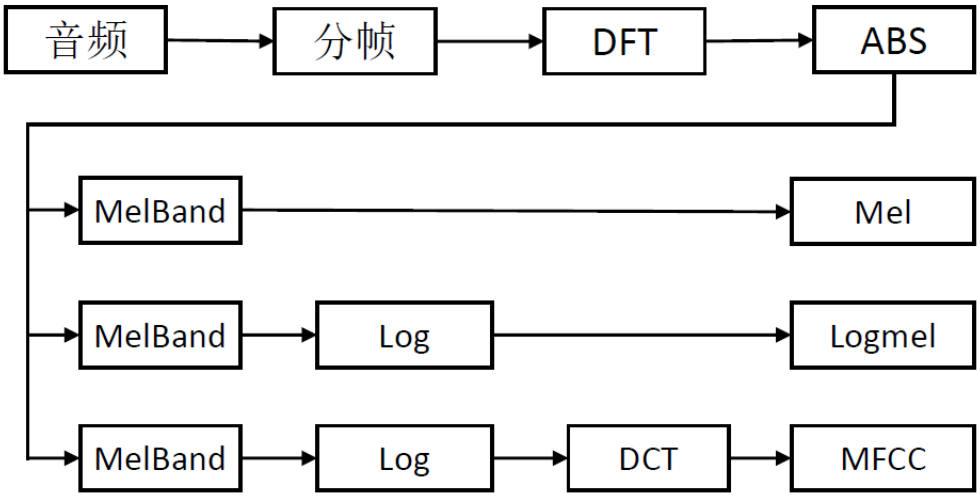

图2 常见频域特征算法框图

根据人的听觉感知特点，不同的声音频率可划分为频带，采用低频线性、高频对数的频带划分方式，例如梅尔频带（Mel bands）。据此可得到上述三种传统特征：先将原始音频数字信号通过（快速）傅里叶变换（DFT、FFT）得到频谱；再将 DFT 的结果按照 Mel bands 划分为 k 个频带，通过每个频带内的频谱能量之相加得到 k 个能量谱作为特征即得到 Mel 谱；对得到的能量谱取对数得到 Logmel 谱；对取对数的结果进行离散余弦变换（DCT）得到倒谱；通常将倒谱的前 13 维作为最终的 MFCC 特征的输出[12]。

2.2关于音频检测的机器学习理论

音频事件检测主要是利用机器学习的模型对音频信号特征进行分类，音乐模型被认为是自动语音识别和音频事件检测在音乐信号上的对应，就像演讲中的鸡尾酒会问题。但这些声音在时间和频率上高度相关，因此有关的神经网络不可直接照搬。而音乐具有与自然语言相似的语法规则或统计规律，这使得自然语言处理的一些技术得以运用。此外因为诸如音符之类的音乐对象可以被识别为时间频

率表示这个二维模式，该技术一定程度上受到了计算机视觉神经网络的影响[4][13]。

该领域传统模型包括最初用于语音识别的高斯混合----隐马尔科夫模型（GMM-HMM）、用于声源分离或音乐转录的非负矩阵分解（NMF）、支持向量机（SVM）、朴素贝叶斯等传统模型[1]。

随着神经网络模型在相关领域的发展，近期音乐事件检测也涌现了一些深度学习模型。Cakir 等人将分帧提取的音频特征输入前馈神经网络（FNN）和卷积神经网络（CNN）分别训练，实验结果在各个实践检测的准确度上都超过了 GMM-HMM[1]。Parascandolo 等人利用双向长短时记忆网络（BLSTM）也有很好的效果[1]。Heittola 等人将 CNN 和循环神经网络（RNN）结合起来得到卷积循环神经网络（CRNN），并在多个数据集上的准确率都超过前述模型[1]。

受到数据集的限制，当前乐器演奏技术检测的研究相对较少，且主要集中于单个音符的演奏技术分类，既有传统的信号处理方法，也有深度学习方法。Wang 等人基于 HMM 给出了竹笛滑音检测算法框架[14]。Chen 等将演奏技法视为时间序列分类问题，给出了电吉他独奏 5 种技法判别的算法框架[15]。Wu 等人研究了合奏中的鼓的 4 种演奏方法的检测[16]。Liang 等对钢琴的延音踏板检测进行了深入研究，并提供了数据集和 CNN 算法框架来分析踏板的存在、起跳和分段[17][18]。Wang 等人录制了 11 种二胡技法的数据集，并利用 FCNN 网络建立了乐句技法判别方法，总体准确率 48%，但在其中 4 种技法上获得了 87%的准确率[19]。

现有的音乐技法鉴别多数涉及的技法种类较少，或能够高效鉴别的技法种类较少，同时缺乏对技法丰富多变的气鸣类乐器的研究。此外，音频事件检测效果较好的 CRNN 模型训练开销和使用时的计算开销都相当大，需要一种可以进行多技法高效判别而计算开销不大的模型。

2.3 常见数据集

Su 和 Chen 等人构建了关于电吉他技法的数据集。数据集通过专业的吉他手记录使用录音接口，对 42 个电吉他音轨进行处理，得到了关于分别带有 6 种技法和空弦音的电吉他音符音频信息。数据集共包括 6,580 个音频片段，涵盖了 11,928 个音符[15][20]。

Wu 等人建立了关于小军鼓的 4 种演奏方法的数据集，包含共计约 13000 个

音频片段[16]。

Liang 等人建立了关于钢琴是否踩下延音踏板的数据集。数据集的基础为 2011 年明尼苏达州国际钢琴比赛上雅马哈 Disklavier 记录的包含踏板信息的 1567 个 MIDI 文件，涵盖了从巴洛克到现代时期的 28 位不同作曲家的作品，并使用经过 Steinway & Sons 认证的软件 Pianoteq 6 PRO3 生成采样率为 44.1 kHz 的音频，音频包括维持踏板信息和删除踏板效果两种，每个音删除开始的 0.5 秒并剔除余下部分不足 0.3 秒的片段，按照 70%、20%、10%的比例划分为训练、验证、测试集[17][18]。

Li 等人建立了基于中国传统民族乐器音频的数据集 A Dataset of Chinese Musical Instruments（DCMI）。该数据集包含了 82 种不同种类或不同调式的中国传统乐器独奏录音，以及各种类型的乐器演奏技巧，每种技法的音频多少不一，多数技法音频不足 3 分钟，且技法通过语音在录音的开始处标注[21]。

第三章 基于机器学习的乐器技法检测

3.1 数据集

该数据集的音频数据是在无消音材料的封闭室内场景下，使用 HUAWEI 荣耀 10 手机内置声卡和驱动软件，对一根 E 调中音竹笛录制而成，录音时保证录音设备到竹笛的距离和角度无明显变化。录音包含 9 种竹笛常见技法和 1 种无技法对应音频。音频为单声道信号，采样率为 48kHz，录音后进行了人工标注，并做了去静音段的操作。

去静音段的方法为采用 liborsa.effects. _signal_to_frame_nonsilent 指示静音帧，并参考了 librosa 开源代码库的其他函数[22]。很多技法，例如音乐中的顿音，本身存在一定量的静音，不同技法静音长度也不尽相同，技法响度与吸气声音响度相对大小也不相同，因此静音判别的阈值按照不同技法人耳听感可分辨为依据进行设置，单吐和滑音设置为 top_db=50, 三吐、双吐设置 top_db=40, 泛音、花舌、颤音、历音、叠音、无技法设置 top_db=30。去掉静音后共计 101 分钟，每种技法的音频时间长度如下表所示：

表 1 各技法去静音后音频长度

技法	颤音	单吐	双吐	三吐	滑音	叠音	历音	泛音	花舌	无
长度	793s	426s	790s	729s	761s	696s	628s	249s	227s	771s

数据集的划分使用分层抽样的方法，用 random.shuffle 函数将数据集按照 6:2:2 的比例随机划分为训练集、测试集和验证集。同一种网络结构先在训练集上训练，再在验证集上评估模型效果，选取测试集上表现最好的模型再在测试集上运行来评估泛化误差。

3.2 音频特征提取

通过 librosa.load 逐秒读入每一种技法 48kHz 的单声道原始音频，并通过 librosa.feature.melspectrogram 计算 mel 谱：先计算 2400 点快速傅里叶变换（FFT），相邻两个窗口间隔 hop_length=375，即每一秒有 128 个计算窗口。对于每一个窗将 FFT 的频谱结果按照 Mel bands 划分为 128 个频带，通过每个频带内的频谱能量之相加得到 128 个能量谱，将其作为特征即得到某种技法这一秒对应的 128*128 的 Mel 谱。

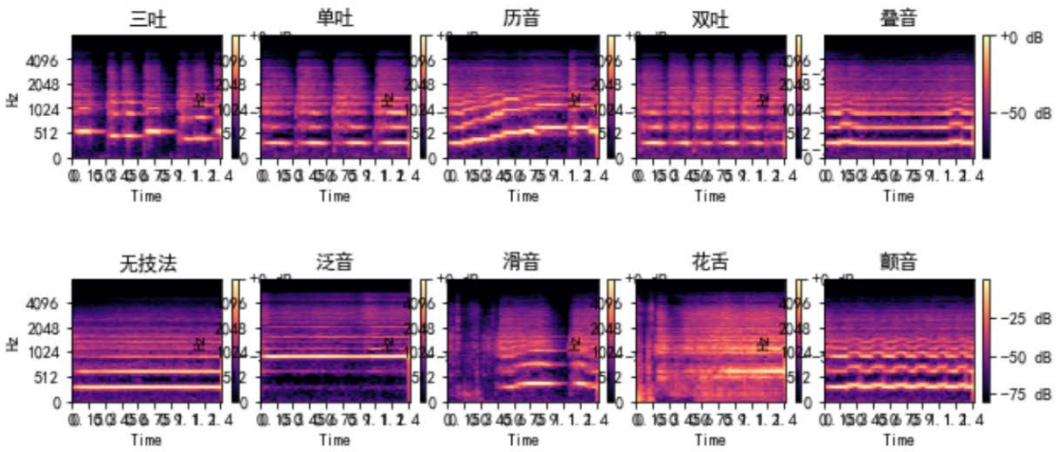

图 3 各种技法音频的 mel-频率能量谱图

通过上图可知，实际训练采用的音频 mel 谱可以作为辨识不同技法的标准，进而作为不同技法音频信号的特征参与训练。

提取特征之后的训练频谱数据有两种，一种为逐秒的 CNN 所需的单个 128*128 频谱数据，另一种为一段乐句对应的多个 128*128 频谱数据。VGG 判别器在训练时每个 batch 的输入为若干个 128*128 的 mel 谱数据，标注为对应的单一技法标注；FCNN 训练时的输入每个 batch 为若干个 10 秒片段对应的 128*1280 的 mel 谱数据，对应 10 秒分别的 10 个技法标注。

3.3 技法检测模型

3.3.1 传统机器学习模型

LeNet 模型是 CNN 网络的经典模型，由 LeCun 在 1998 年提出用于解决邮政编码识别问题[23]。网络结构如下图所示，图中浅蓝色为输入层，蓝绿色为不补零的卷积层，黄色为 ReLU 层，浅黄色为最大值池化层，最后三层为全连接层。

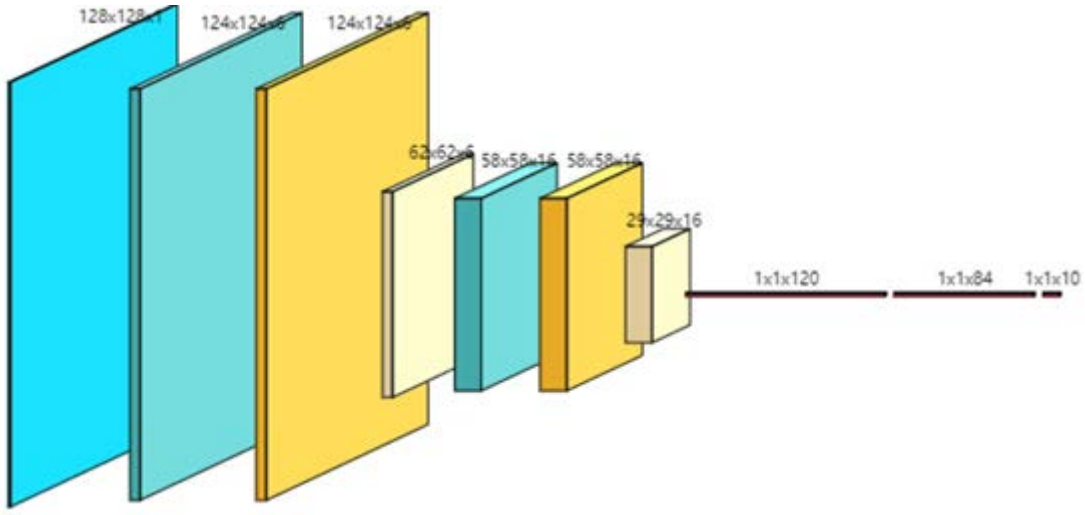

图3 LeNet-5网络结构框图

此处将输入结构调整为 128*128 的 mel 谱图，层数、各层卷积层尺寸与数量、降采样层的参数跟 LeNet-5 均相同，激活函数亦采用最传统的 sigmoid 函数。由于输入尺寸不同，第一个全连接层的参数数量也不同，余下均相同。

3.3.2 VGG 模型

VGG 模型是牛津大学 Visual Geometry Group 于 2014 年提出的神经网络，该网络证明了增加网络的深度能够在一定程度上影响网络在图片识别任务上的的最终性能[24]。这里选用 VGG13 和 VGG16 作为网络结构，用于训练不同技法的时频 mel 谱图的识别，得到一个逐秒技法检测判别器。

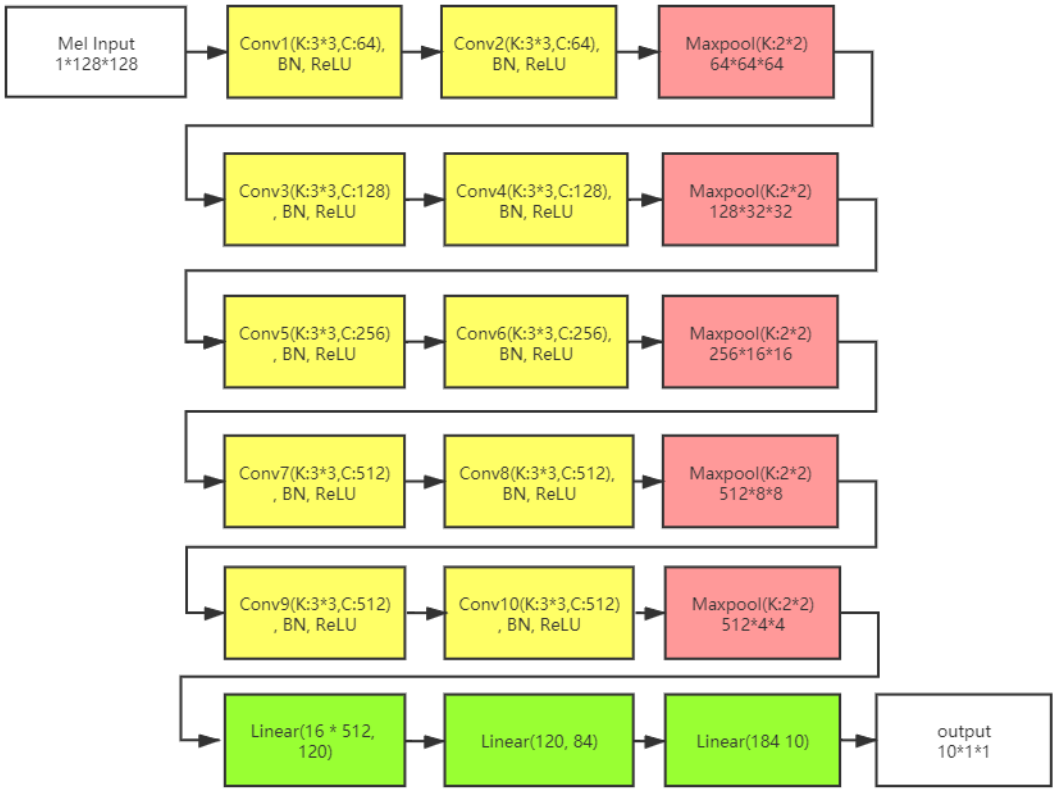

图 4 VGG13 网络结构框图

VGG13 由 5 组“卷积，卷积，池化”和三层全连接依次连接而成。相邻的两个卷积层的卷积核尺寸均为 3*3，在保证感受野相同的情况下大大节约了参数数量，同时多层非线性层也增加了网络深度，为习得更复杂的模式提供了可能。卷积之后通过 Batch Norm 层保持在该层输入输出的数据分布相同或相近，之后通过一个 ReLU 层再进入下一层卷积。池化层为最大值池化，步长和 kernel 的大小均为 2。最后的全连接层与 LeNet-5 最后的全连接层尺寸相似。

模型的损失函数为多分类交叉熵损失 CrossEntropyLoss。训练算法为随机梯度下降法，学习率为 0.001，动量为 0.9。

VGG16 由两组“卷积，卷积，池化”和三组“卷积，卷积，卷积，池化”依次串联得到，ReLU、BatchNorm 等结构和训练算法等参数均与 VGG13 相同。

3.3.3 FCNN 模型

全卷积神经网络（FCNN）是在 CNN 的基础上增加“转置卷积”层得到的一

种深度神经网络。FCNN 最初应用于计算机世界的图片分割，网络的输出尺寸和池化尺寸可以根据输入特征图片大小自适应地调整[10]。因此可以利用 FCNN 的特点构建不定长度输入的技法判别器，可以避免传统算法因尺寸归一化引发的精读损失。虽然输入尺寸为 10*mel 谱，但训练结束后保证网络结构和参数不变，改变输入的时间维上的尺寸，输出也会随之增加或减少相同的倍数，最终输出的技法判别序列的项数和输入的视频 mel 谱序列的项数相一致。FCNN 的网络结构如下图所示：

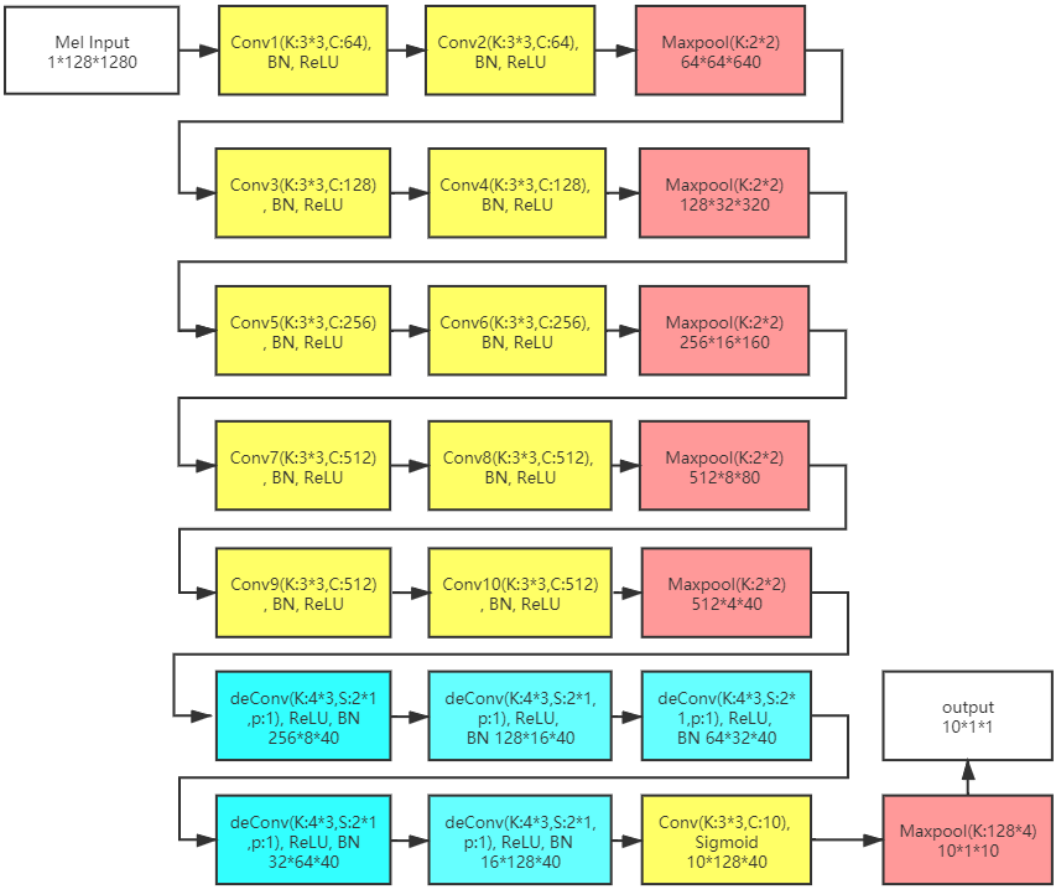

图 5 基于 VGG-13 的 FCNN 网络结构框图

对于每一个 batch，FCNN 的输入参数为 1*128*(128*t)，首先经过上一小节所述 VGG13 全部卷积层和池化层的计算，每个 batch 得到 512*4*(4*t)，其中 512 为通道数。之后不经过全连接，而是直接经过 5 个相连的卷积核大小为 4*3、步长为 2*1、padding 为 1 的转置卷积层，并依次经过 ReLU 层和 BatchNorm 层。

最终，得到的输出为 batch size*16*128*(4*t)，频域的 128 个声带特征全部恢复，事实上有些技法可能只在某些声带中才能得以体现。之后经过 10 个卷积核和 sigmoid 激活函数，和 kernel 尺寸为 128*4 的最大值池化层，每一组 batch 都可以得到 10 个 1*t 的判别结果（10 为技法可能的种类数），作为 t 秒音频技法的预测概率。

FCNN 的损失函数和训练算法跟 VGG13、VGG16 相同。

3.4 实验结果

利用第三章建立的数据集中的训练集训练数据，训练 4 组 VGG13、4 组 VGG16 和 8 组 FCNN，每种网络结构选取在验证集上表现最好的模型在测试集上评估泛化结果。实验结果如下表所示：

表 2 不同模型的验证集技法平均准确率

模型	N-Bayes	LeNet-5	VGG13	VGG16	FCNN
准确率	23-25%	70-74%	87-90%	87-89%	88-94%

通过实验可以看出神经网络模型均远远高于传统机器学习算法。AddBoost、随机森林、高斯朴素 Bayes、线性回归等方法准确率均低于参数为 1.0 的多项式朴素贝叶斯法，且 PCA 降维后结果没有明显改善。神经网络模型当中 VGG 的效果要明显优于 LeNet-5，而且部分 FCNN 的效果优于各组 VGG13 和 VGG16，并且 VGG13 和 VGG16 性能无明显差异说明 FCNN 相对 VGG13 的优势并非由单纯的层数增加导致的，而是由转置卷积层的性质导致的。

表 3 不同模型验证集上各类技法准确率

	LeNet-5	VGG13-0	VGG13-1	FCNN-5	FCNN-4
总体	73%	87%	88%	93%	93%
无技法	76%	84%	85%	93%	93%
单吐	56%	84%	90%	91%	94%
双吐	75%	80%	90%	95%	92%

三吐	65%	88%	91%	89%	91%
颤音	87%	96%	96%	95%	96%
滑音	62%	84%	82%	92%	86%
叠音	73%	94%	96%	95%	94%
历音	88%	90%	83%	99%	96%
泛音	68%	78%	78%	100%	97%
花舌	66%	83%	77%	68%	84%

我们选取在验证集上表现最好的 LeNet-5，表现最好的两个 VGG（编号分别为 0、1）和表现最好的两个 FCNN（编号分别为 4、5）。结合以上实验结果可知，除了 LeNet-5 准确率较高的历音以外，VGG13 明显优于 LeNet-5。而 FCNN 在无技法、单吐、双吐、滑音、历音和泛音这 6 个类别的音频上准确率均明显优于 VGG13；此外，VGG 准确率相当高的颤音、叠音、历音和三吐当中，除了 FCNN 判别准确率接近 100%的历音以外，其余三项 VGG13 和 FCNN 的结果均无明显差异，可能是由于 VGG 提升空间不大所致；而 FCNN 预测的花舌准确率相较 VGG13 在某些训练初始值时有明显下降，个别初值跟 LeNet-5 无明显差异，具有更大的不稳定性，可能是由于数据集中花舌技法的音频数据较少，并且未采取数据增强手段，单项准确率对整体影响较小导致。总体可知，增加转置卷积层的 FCNN 在判别性功能上显著优于 VGG。

表 4 不同输入时长的验证集、测试集准确率

输入长度	5s	10s	20s	50s
FCNN-4	94%, 92%	94%, 93%	94%, 92%	93%, 93%
FCNN-5	94%, 92%	94%, 93%	94%, 93%	94%, 94%

我们将数据集中的验证集和测试集随机打乱，拼接成不同长度的乐器音频输入到模型评估准确率，实验结果入上表所示，前一个数字为验证集上准确率，后一个数字为测试集上准确率。实验结果说明，输入音频的长度对于竹笛技法判别的准确率没有明显影响，故而 FCNN 模型可用于鉴别不同时长的输入音频中的

竹笛技法检测。

此外，在 DCMI[21]长度共计约 3 分钟的 G 调新笛技法数据集上，用表现最好的 FCNN-4 和 FCNN-5 做测试，准确率均不高于 20%，考虑到本数据集和 DCMI 各自数据集内部在噪音、笛子调式等性质上无明显差异，造成该现象有如下可能原因：技法使用不熟练、不清晰导致技法客观听感上容易混淆，例如历音在不清晰时可能被误认为或误演奏为滑音；技法使用时包含了旋律及旋律变化信息，音频录制时三吐和叠音包含了旋律变化信息，DCMI 的无技法音阶和叠音可能因旋律变化被误判为三吐，很多其他技法也被误判为叠音；音频内技法使用密集程度有关，音频录制处理时，叠音技法频率相对密集，跟颤音有一定的相似性，进而造成了大量误判。后续研究可以增加各种变化旋律下的技法数据并增加 dropout 层防止过拟合。

第四章 总结与展望

本文构建了关于中国竹笛 10 种技法的音频数据集，数据量相比较以往数据扩大了很多。并且本文以竹笛为例，提出了一种基于 FCNN 网络结构的不定时长输入、端到端的音频事件的检测方法用于，乐器技法检测。我们针对固定长度的音频建模，构建了该竹笛技法判别器，检验了该模型在不同时长输入下的判别稳定性，并比较了该模型对各类技法的判别正确率，证明了其相对于 VGG13 和其他模型的优越性。

该模型还有很多地方值得后续研究。模型那你在 DCMI 数据集上的表现说明该数据集还有很多地方值得完善：应增加技法相对旋律及旋律变化的丰富性，在各种相同的旋律上录制相同的技法，并尽量保证各种技法的数据量相近。同时在训练模型时，每一个训练单元可以从 1s 缩短到 0.25s 到 0.5s，理论上包含技法的几十个周期即可作为判别依据，不必包含 1s 内的全部信息，同时实用领域 1s 的时长也过长，在节奏较快的乐曲中技法事件的精确度为 1s 相对模糊，不够精确。

其次，模型没有在真实世界所录的乐曲中检测技法判定准确率，本文数据集中的音频计算 mel 谱时以秒为单位，时间窗在这一秒中的任何位置均为所标注的技法对应的音频，但实际的乐曲演奏录音中，可能包含单一的远远短于一秒的技法使用，在某种意义上为对该一秒音频的弱标注，因而需要在数据集中加入真实演奏的录音或技法拼接无技法的复合音频。

此外，实际录音中可能包含不同的噪声，模型对于不同强度的噪音变化是否有鲁棒性还有待探索，后续的模型可以尝试在提取特征时采用基于降噪解码器的特征进行训练，并在加各类白噪声、无技法音频、随机置零等情况下的鲁棒性。在上述更复杂的数据集或更复杂的噪音等情况中增加 dropout 层防止对噪声的过拟合，评估其效果。

很多时候层数较浅的 VGG 可能没有很好的效果，或训练使用的时空开销较大。用更新的 ResNet 代替 VGG13 作为基线，用 R-FCNN 代替基于 VGG 的 FCNN 是潜在的可行途径。

本文工作对未来很多工作具有重要意义。此项工作有助于未来进一步的音乐

对象识别和记录，以实现完整的自动音乐转录，这不仅包括提取音符音高，还包括音符相关的技法检测。此外，这项工作还对鉴别竹笛音乐及其他声乐或器乐的流派、版本鉴别，有关口传文化和有声文化的记录、保护和研究，民族音乐及中国有声传统文化的学习、教育与传承有一定的指导意义。

参考文献

[1] Adavanne, Sharath, et al. *Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features.* (2017).
[2]A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription.* New York: Springer, 2006.
[3] E. Benetos, S. Dixon, D. Giannoulis, et al. “Automatic music transcription: challenges and future directions,” *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, Dec. 2013.
[4] Benetos, E. , Dixon, S. , Duan, Z. , et al. Automatic music transcription: an overview. *IEEE Signal Processing Magazine*, 36(1), 20-30. 2019.
[5]M. Müller, D. P. Ellis, A. Klapuri, et al. “Signal processing for music analysis,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, Oct. 2011.
[6] M. Schedl, E. Gómez, and J. Urbano, “Music information retrieval: Recent developments and applications,” *Foundations and Trends in Information Retrieval*, vol. 8, pp. 127–261, 2014.
[7] FJ Ca˜nadas Quesada, N Ruiz Reyes, et al. “A multiple-f0 estimation approach based on gaussian spectral modelling for polyphonic music transcription,” *Journal of New Music Research*, vol. 39, no. 1, pp. 93–107, 2010.
[8] Fabrizio Argenti, Paolo Nesi, and Gianni Pantaleo, “Automatic transcription of polyphonic music based on the constant-q bispectral analysis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1610–1630, 2010.
[9] 袁静芳. *民族器乐(修订版)*. 北京. 高等教育出版社, 2004. pp. 43–47.
[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431– 3440.
[11] 王开武. 基于深度神经网络的异常声音事件检测. 重庆: 重庆大学光电工程学院, 2018.
[12] 周建超. 基于深度学习的鲁棒重叠音频事件检测研究. 北京: 北京大学信息科学技术学院, 2018.
[13] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, “Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription,” in *Proc. International Conference on Machine Learning* (ICML), 2012.
[14] C. Wang, E. Benetos, X. Meng, E. Chew. Towards HMM-based glissando detection for recordings of Chinese bamboo flute. *International Society for Music Information Retrieval Conference Late-Breaking Demos Session*, 2018.
[15] Yuan-Ping Chen, Li Su, Yi-Hsuan Yang, et al., “Electric guitar playing technique detection in real-world recording based on f0 sequence pattern recognition.,” in *ISMIR*, 2015, pp. 708–714
[16] Chih-Wei Wu and Alexander Lerch, “On drum playing technique detection in polyphonic mixtures.,” in *ISMIR*, 2016, pp. 218–224.

[17] Beici Liang, Gy¨orgy Fazekas, and Mark Sandler, “Piano sustain-pedal detection using convolutional neural networks,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 241–245
[18] Beici Liang, Gy¨orgy Fazekas, and Mark Sandler, “Towards the detection of piano pedalling techniques from audio signal,” . 2017.
[19]Z. Wang, J. Li, X. Chen, et al. Musical Instrument Playing Technique Detection Based on FCN: Using Chinese Bowed-Stringed Instrument as an Example. arxiv, 2019.
[20] Li Su, Li-Fan Yu, and Yi-Hsuan Yang, “Sparse cepstral, phase codes for guitar playing technique classification.,” in *ISMIR*, 2014, pp. 9–14.
[21] Z. Li, X. Liang, J. Liu, et al. “DCMI: A database of chinese musical instruments,”. *The Conference on Sound and Music Technology*. 2018.
[22] Mcfee B , Raffel C , Liang D , et al. librosa: Audio and Music Signal Analysis in Python. *Python in Science Conference*. 2015.
[23] Lecun, Y , and L. Bottou . "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11(1998):P.2278-2324.
[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

本科期间的主要工作和成果

本科期间参加的主要科研项目

1. 随机矩阵基本理论. 本科生科研项目.
2. 语音通信系统仿真. 课程项目.
3. 中国民族复调音乐乐器检测. 本科生科研项目.
4. 中国民族器乐技法检测. 毕业设计项目.
5. 西南官话音调语调与四川民歌旋律的关联. 暑期科研项目.
6. 粤语歌曲基频与歌词声调的关联. 课程项目.

本研基金

1. 基础数学拔尖人才计划. 教育部人才培养专项. 李若. 2018,2019.

期刊：

无

会议论文：

1. 李子晋, 蒋超亚, 陈晓鸥, 马英浩, 韩宝强. 基于卷积循环神经网的复音音乐中国民族乐器检测. 全国声音和音乐技术会议, 中国哈尔滨, 2019 年 12 月.

专利：

无

致谢

我要感谢北京大学王选计算机研究所的数字音频信息处理实验室。感谢课题组的陈晓鸥研究员，他在我刚进入音乐信息检索领域时提供了很多科研的指导和机会，并对毕业论文提供了很多帮助和支持。课题组的蒋超亚师李婧如同学就本文的编程工作提供了可供参考的代码和必要的帮助，在此向他们表达感谢。同时，我也要感谢课题组的其他老师、同学和毕业生们，和他们的讨论交流使我受益良多。此外，本文代码还受益于胡俊峰副教授的《Python 程序设计与数据挖掘》，在此向胡老师、课程助教和共同探讨作业的同学表达感谢。

我要特别向相关领域的两位老前辈表达感谢。爱丁堡大学 Nigel Osborne 院士是（计算机）作曲、音乐感知、音乐信息检索和音乐治疗领域的巨擘。同时，他通晓 10 国外语，是世界音乐的专家，有极强的国际视野和同理心。他还胸怀天下、知行合一，是叙利亚儿童音乐治疗心理康复的负责人。Nigel 谦逊、儒雅、风趣、平和，冒生命危险挽救他人，学艺双馨、言行如一，热心帮助晚辈，在 2019 年秋天到访北大时让我明白音乐信息检索、语义理解与人对音乐的感知、理解的辩证关系，他的鼓励和拥抱给我极大的帮助和鼓舞。他树立了一个难以超越的学者典范，激励青年学者努力奋斗。另一位是中国音乐学院韩宝强教授，他是国内音乐科技领域的老前辈，在 2019 年秋天的北大艺术学院研讨会上有很多富有洞察性的观点，提出了很多有待解决的问题，涉及领域广泛，包括美学、乐器学、音乐文化生态、音乐教育、乐器演奏、音乐认知等，令我很受启发，为青年学者树立了前进方向。在此向两位老前辈表达诚挚的感谢。

我还要感谢北京大学数学科学学院的王杰教授和艺术学院毕明辉副教授。他们的课程给本人很大启发，王杰老师的《音乐与数学》对本人后续科研和升学提供了极大帮助，毕明辉老师学识渊博、治学严谨、中西贯通，投入极大精力教学和帮助学生，他的的《中国民族器乐经典》和《西方音乐欣赏》观点深刻、论据详实、极具启发意义，直接促成本人转向（中国）音乐信息检索领域，并对本文的研究齐了一定指导作用。再次向两位北大授课教师表达感谢。

我还要感谢北京大学（学生）中国音乐学社。社员之间相互的帮助、支持和期待是我迷茫生活中的一盏明灯，“爱校、弘乐、感恩、合作、正己、育人”的社训让我受益匪浅，社团的挑战成为我前进的动力。此外，和社员丁明朔同学讨论神经网络等算法让我受益良多，社团毕业生上海纽约大学助理教授夏光宇、北京邮电大学李荣锋老师和卡内基梅隆大学在读博士生戴舒琪师姐为我转入该研究领域提供了帮助和支持。提供支持的还有帅气的字节跳动工程师孔秋强博士，罗切斯特大学段智尧助理教授及其音频信息实验室，伦敦玛丽女王大学 Emmanouil Benetos，数学科学学院李知含等人。在此向他们表示感谢。

此外，我的父母在疫情期间为研究提供后勤保障；青千班主任刘保平向我展示了青年学者的真实状态；学院冯美娜老师在党团、学工方面的帮助；室友和我共同营造的良好氛围，特别是高乾同学和我打赌来督促我早起；转变科研领域、升学和疫情期间，包括不局限于胡明源、夏华钦、陈肇盟、苏文杰等挚友提供了必要的鼓励和支持；北大钢琴社多任社长、提琴社首席和音乐创作协会会长和校友等人亦提供了帮助和支持。在此一并表达感谢。

最后，令人愉悦的作品被认为能提升人的短期创造力，为疫情期间的研究提供了帮助。《贝多芬第 5 交响曲》激起共鸣和斗志；听《门德尔松 e 小调小提琴协奏曲》和王丹红的《太阳颂》如同嗑药，欲罢不能；郭文景《滇西土风》《戏》对音乐的创造性理解巧夺天工；赵季平颇有法国遗风的色彩感音乐扣人心弦；关乃忠《龙年新世纪》双打击协奏曲妙不可言；香港中乐团艺术总监兼首席指挥阎惠昌的造诣高超；深刻理解音色的指挥顾宝文让我对著名青年作曲家李博禅的优秀作品《楚颂》的评价从“几乎处处不好听”变成“woc 真厉害”；还有彭修文、卢亮辉、苏文庆几位前辈也令我受益匪浅；北大中文系孔江平教授采风时记录的丰富的口传文化及其深入的研究让我叹为观止；B 站 up 主观视频工作室上传的带有清晰语义和鲜明立场的音频信号让我醍醐灌顶、顿觉荡气回肠；Fb 主页 ClassicFM、华乐谜团和北大中乐学社也提供了帮助。在此向有关工作者们表达感谢。