基于机器学习的中国传统乐器音频识别

摘　要

随着社会的快速发展以及信息化时代的来临，音乐信息检索技术开始广泛地应用在我们的生活中。几乎所有的音乐数据都有与之对应的乐器，因此研究乐器分类对音乐信息检索有着非常重要的意义。通过声音来识别乐器，是通过人工智能研究中国传统乐器的各种课题的前提。目前的乐器音频识别并没有专门针对中国传统乐器的，因此我们要开展这项有关中国传统乐器音频识别的研究。

本文首先介绍了乐器音频识别的理论基础，包括音频信号的理论基础和音频识别的整个流程。然后自己搭建了中国传统乐器音频数据库，包括 78 件乐器和 1630 个音频文件。接着本文研究了中国传统乐器音频的乐器种类识别问题，利用梅尔频谱特征作为输入，训练了 8 层的卷积神经网络，最终取得了 99.3%的准确率，随后从三个角度对实验结果进行了分析。接着本文研究了中国传统乐器音频的演奏技巧识别问题，首先进行了单一乐器的演奏技巧识别，利用预训练的 ResNet 模型提取梅尔频谱的特征，然后通过 SVM 分类在所有乐器上均取得了 99%的准确率，接着为了提高模型的泛化性，提出了同一类乐器的演奏技巧识别，通过训练卷积神经网络进行实现，最终四大类乐器的识别准确率如下：吹奏类乐器 95.7%，弹拨类乐器 82.2%，拉弦类乐器 88.3%，敲击类乐器 97.5%。

我们开放了中国传统乐器音频数据库和整个实验的 Python 源代码以供进一步研究。

关键词　乐器识别　演奏技巧　特征提取　机器学习

Audio Recognition of Chinese Traditional Instruments based on Machine Learning

ABSTRACT

With the rapid development of society and the advent of the information age, music information retrieval technology has been widely used in our lives. Almost all music data have corresponding Musical Instruments, so the study of musical instrument classification is of great significance for music information retrieval. Using audio to identify Musical Instruments is the premise of studying various subjects of traditional Chinese Musical Instruments through artificial intelligence. At present, the audio recognition of Musical Instruments is not specific to traditional Chinese Musical Instruments, so we want to carry out this research on the audio recognition of traditional Chinese Musical Instruments.

This paper first introduces the theoretical basis of musical instrument audio recognition, including the theoretical basis of audio signal and the whole process of audio recognition. Then I built the audio database of traditional Chinese Musical Instruments, including 78 Musical Instruments and 1630 audio files. This paper studies the problem of musical instrument type identification of traditional Chinese musical instrument audio, trains the 8-layer convolutional neural network with MEL spectrum feature as input, and finally achieves 99.3% accuracy, and then analyzes the experimental results from three angles. After that, this paper studies the playing technique identification of Chinese traditional musical instruments. First, the playing technique identification of a single instrument is carried out, and the characteristics of Mel's spectrum are extracted by using the pre-trained ResNet model, then SVM is used to classify all musical instruments with an accuracy of 99%.Then, in order to improve the generalization of the model, the paper proposes the recognition of playing technique of the same kind of instruments, which is realized by training the convolutional neural network. Finally, the recognition accuracy of the four kinds of instruments is as follows: 95.7% for blowing instruments, 82.2% for plucking instruments, 88.3% for pulling strings, and 97.5% for percussion instruments.

We open source the audio database of traditional Chinese Musical Instruments and the Python source code of the whole experiment for further research.

KEY WORDS instrument recognition playing technique feature extraction machine learning

目　录

第一章　引言

1.1　研究背景及意义

中国是一个多民族的国家，传统乐器的种类繁多。中国的传统乐器历史悠久，源远流长。根据其演奏方式不同可分为四大类：吹奏类，如笛，笙，埙，箫，巴乌，吐良等；敲击类，如木鱼，梆子，鼓，编钟，锣，镲，铙钹等；弹拨类，如古筝，琵琶，中阮，扬琴，三弦，筚篥等；拉弦类，如二胡，板胡，二弦，中胡，椰胡等。

随着社会的快速发展以及信息化时代的来临，越来越多的音频数据出现在了我们的生活中。怎样快速有效地从大量的音频数据中找到自己感兴趣的数据，并将之合理化地管理成为了人们研究的热点。随着人工智能技术的逐渐成熟，音乐信息检索（Music Information Retrieval，MIR）技术开始广泛地应用在了我们的生活中。几乎所有的音乐数据都有与之对应的乐器，因此研究乐器分类对音乐信息检索有着非常重要的意义。通过声音识别乐器的种类是通过人工智能研究中国传统乐器的各种课题的前提工作，利用乐器识别，可以完成乐器音频自动分类，乐器音频自动检索，乐器音频自动作曲等工作。目前的乐器识别并没有专门针对中国乐器的，因此我们要开展这项有关中国传统乐器音频识别的研究。

中国传统乐器因其形制和物理声学的特点，具有有别于西洋乐器的演奏技法及音频特征，中国传统乐器作为世界乐器的重要组成部分，具有极大的研究价值，而音频识别与分类又是数字音频领域的经典问题，因此研究中国传统乐器音频识别在科学研究和实际应用中具有深刻的意义。特征提取是实现乐器分类的关键，由于机器学习在图像处理和自然语言理解等方面具有显著的特征提取优势，所以选择机器学习来解决中国传统乐器音频的识别问题。我们主要研究基于机器学习的中国传统乐器音频识别问题。

1.2　国内外发展现状

本文主要是研究中国传统乐器音频的识别问题，属于音频分类及识别的范畴。而对于音频的分类及识别，影响识别准确率的关键在于音频特征的提取。

早在 2002 年，P.Cook 等人[1]在对音乐类型进行分类时，提出了一种表示音色信息的特征组合。所选取的主要特征是频谱质心的均值和方差、频谱衰减的均值和方差、频谱通量的均值和方差、过零率的均值和方差、低能量值（检测窗口的 RMS 能量低于整个窗口的 RMS 能量的百分比）和梅尔频率倒谱系数(MFCC)的前 5 个系数的均值和方差，共 19 维特征向量。他们对古典音乐数据集和爵士音乐数据集进行了分类，分别取得了 64%和 57%的准确率。P.Cook 等人[1]在进行实验时，虽然所选取的特征众多，但是并未考虑每种特征对于分类的影响，可能某些特征对于音频分类识别的贡献度很小甚至是冗余信息。2008 年 C.Simmermacher 等人[2]提出在实际应用中，常用的特征提取方案内部存在冗余，发现一组紧凑且有效的特征集是解决问题的关键。他们主要考虑了三种类型的特征：基于人类感知的特征，MFCC 特征和 MPEG-7 音频描述符，如表 1-1 所示。最后发现 MFCC 和基于人类感知的特征在乐器分类中占主导地位，MFCC 特征提供了最佳的分类性能。最后他们对 20 种乐器进行分类达到了 86.9%的准确率。2010 年 J. Liu

等人[3]对 8 个家族的乐器进行分类，提取了短时傅里叶变换（STFT），MFCC，频谱峰值因数（SCF）和频谱平坦度测量（SFM）特征，最终发现 MFCC 的性能最佳，训练集和测试集的平均准确率达到 91.34%。因此，我们在进行乐器分类时会首先考虑 MFCC 特征。

表 1-1 常见的音频特征提取方案

编号	描述	类别
1	零交叉	基于人类感知
2-3	过零率均值和方差	
4-5	均方根均值和方差	
6-7	频谱质心均值和方差	
8-9	频谱带宽均值和方差	
10-11	频谱通量均值和方差	
12	谐波质心	MPEG-7 音色描述符
13	谐波偏差	
14	谐波扩展	
15	谐波变化	
16	频谱质心	
17	对数起音时间	
18	时间质心	
19-44	MFCC 的前 13 个系数的均值和方差	MFCC

M.S.Nagawade 等人[4]利用 MFCC 提取特征，使用 KNN 进行识别，最终在大提琴，钢琴和小号上取得了 91.66％的识别准确度，在长笛和小提琴上取得了 83.33％的识别准确度。他们发现，虽然 MFCC 特征的识别精度较高，但是随着分类乐器数量的增加，识别精度开始下降。因此需要选择一种可以区分更多信息的分类器。2003 年，N.C.Maddage 等人[5]利用 MFCC，频谱功率，过零率等特征对音频进行分类时发现，SVM 分类器的效果要优于 K 近邻模型，高斯混合模型（GMM）和隐马尔可夫模型（HMM）。SVM 分类器的错误率为 6.86%，而 KNN，GMM 和 HMM 的错误率分别为 20.57%，12.31%和 11.94%。2006 年，M.Ogihara 等人[6]在进行音频分类时，选用了高斯混合模型，k 近邻模型，线性判别模型（LDA）和支持向量机等分类器，最终发现支持向量机（SVM）的识别效果最佳。2008 年，C.Simmermacher 等人[2]提取了包括 MFCC 在内的 44 维特征进行乐器分类，使用了 KNN，贝叶斯，SVM，MLP，RBF 等模型，最终其性能如下：KNN95.75%，Bayes86.5%，SVM97.0%，MLP95.25%，RBF95.0%，其中 SVM 取得了最佳的效果。因此对于传统的机器学习方法我们一般使用 SVM 作为分类器。

由上文可知，对于传统的乐器音频识别问题，我们选择使用 MFCC 特征和 SVM 分类器。随着深度学习的不断发展，人们开始发现神经网络在高维特征分类时的优势，在使用大量数据进行神经网络的训练时可以取得比传统机器学习方法更优秀的性能。2015 年 S.Masood 等人[7]使用了包含 MFCC 在内的 20 维特征向量作为输入，使用人工神经网络固定输入帧进行训练，最终取得了 89.17%的平均准确率。2018 年王飞[8]使用听觉谱图特征作为输入，分别利用卷积神经网络，全连接神经网络和 SVM 进行乐器识别，最

终取得了 96.9%，95.7%，88.9%的准确率。2019 年，Hendrik Purwins [9]等人提到 MFCC 特征一直被用作音频分析任务的主要声学特征表示，利用 DCT 变换得到 MFCC 时破坏了频谱图的空间信息，因此在深度学习中我们一般选用 DCT 变换之前的梅尔频谱图作为提取的特征。如表 1-2 所示，对于 MFCC 特征，将其展平我们可以输入到全连接神经网络中进行训练，而对于梅尔频谱特征，我们可以直接将其输入到卷积神经网络中进行训练，既保留了时域信息，同时也提取了频域的音色信息。因此对于目前的乐器音频识别问题，我们一般使用梅尔频谱特征和卷积神经网络进行识别，当数据量过小时，我们会选择使用 SVM 分类器。

表 1-2 常见的神经网络模型关于音频序列的应用

神经网络	输入	特点
全连接神经网络	一维向量	学习目标的全局特征
卷积神经网络	一维向量（1D-CNN） 二维频谱（2D-CNN）	学习目标的局部特征
循环神经网络	多个一维向量组合	学习长短时间相关性

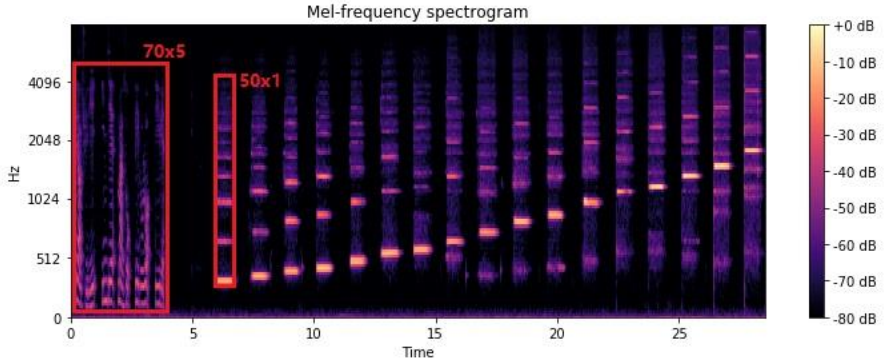

图 1-1 不同大小滤波器的提取信息能力

2017 年 J.Pons 等人[10]利用卷积神经网络对音乐的音频信号进行音色的分析时，认为首层滤波器的形状对于提取音高和时间信息来说非常重要。为了对音色信息进行充分的提取，就必须对振幅频谱进行有效的滤波。如图 1-1 所示，只有滤波器捕获整个频谱包络，才能有效表达音色信息。据此他们提出了一种卷积神经网络的设计思路，来满足音高不变性，响度不变性，持续时间不变性和空间位置不变性四大属性。他们分别设计了两种不同的架构对 11 种乐器进行分类，单层架构使用较窄的滤波器（m*1）进行单个音色的具体识别，多层架构使用较小的滤波器（3*3）进行具体音色的小范围识别。虽然保持了音高不变性，但是由于网络层数较小，导致学习能力有限。本文在使用 3*3 的小滤波器的基础上，搭建了 8 层的卷积神经网络，提高了可学习的参数数量，同时通过调整超参数减少了过拟合的程度。2018 年，Qiuqiang Kong 等人[11]在参加 2018DCASE 挑战赛时，分别基于 VGGNet 和 AlexNet 搭建了 8 层和 4 层的卷积神经网络，使用对数梅尔频谱作为输入，在音频分类识别中取得了 92.8%和 89.5%的准确率。具体的卷积神经网络如表 1-3 所示。他们在进行特征输入时，选择将特征进行切割，然后重组成相同维度以满足输入要求，这对数据集的质量要求很高。本文与之相比，在进行预处理前进

行了音频数据的降噪和切割，减小了实验误差。具体的卷积神经网络的识别方案会在第三章中进行阐述，第三章主要使用梅尔频谱特征，训练卷积神经网络来进行识别工作。

表 1-3 CNN4 与 CNN8 的具体架构

Feature size	CNN4	CNN8
64	Log Mel Spectrogram	
32	5*5*64	3*3*64 3*3*64
	2*2, max pooling	
16	5*5*128	3*3*128 3*3*128
	2*2, max pooling	
8	5*5*256	3*3*256 3*3*256
	2*2, max pooling	
4	5*5*512	3*3*512 3*3*512
	2*2, max pooling	
1	Global max pooling	
	Sigmoid or SoftMax	
Parameters	4,309,450	4,691,274

国内最早也进行过有关音频识别方面的研究，也都取得了不错的进展。但是基本所有研究都是针对西方乐器数据库的，很少有人专门针对中国传统乐器进行研究。2010 年 J. Liu 等人[3]使用 MFCC 特征，通过支持向量机（SVM）对 13 种中国乐器和 13 种西方乐器进行分类，测试集准确率达到 87.23%。2012 年沈骏等人[12]在 MPEG-7 特征的基础上加入自己设定的新特征，利用 K 近邻算法，在二胡，扬琴，鼓等 9 件乐器上取得了 76.45%的置信度。2017 年，王芳[13]自己搭建了包含 6 种中国乐器的数据库。使用了标准 MFCC 以及其一阶、二阶差分参数。通过深度置信网络对其进行分类，取得了 99.2%的准确率。虽然他们的模型乐器识别准确率较高，但是他们所使用的乐器音频数据库过小，而且所涉及乐器数量较少，无法代表中国传统乐器这个整体，仅可作为实验性质，不具有较强的鲁棒性。因此本文中会搭建一个包含 78 件中国传统乐器的音频数据库以供实验和研究。表 1-4 为本文中具体使用的中国传统乐器音频数据库与他人实验所使用的数据库的对比。具体有关中国传统乐器音频数据库的搭建会在第二章中进行阐述。

表 1-4 中国传统乐器数据库对比

研究者	乐器数量	数据集大小
J. Liu	13	2177
沈骏	9	2700
王芳	6	600
Our	78	18819

由于中国传统乐器的数量较多，同时每种乐器的演奏技巧复杂，同一件乐器既可以充当弹拨乐器，也可以作为打击乐器，单纯的音色分析无法描述乐器本身，而且目前国内外并没有专门针对乐器的演奏技巧的研究，因此在本文中要进行有关中国传统乐器音频的演奏技巧识别的研究。演奏技巧识别和乐器种类的识别，本质上是相同的，都是提取梅尔频谱特征进行识别。不同之处在于，对于乐器种类的识别，每种乐器的数据量较为充足，适合使用神经网络进行训练识别。而对于乐器演奏技巧的识别，每种乐器的演奏技巧的数据量较少，所以选择使用预训练的神经网络提取梅尔频谱的特征，然后利用 SVM 进行识别。具体实验将会在第四章中进行详细叙述，同时在第四章中还提出了使用同一大类乐器进行演奏技巧的识别，在增加数据量之后就可以训练卷积神经网络来进行演奏技巧识别了。

1.3 研究内容和目标

本文的研究对象是中国传统乐器，研究目标主要是对中国传统乐器音频的乐器种类和演奏技巧两方面进行识别。在研究过程中，首先自己构建了用于训练的音频数据库，然后从时域和频域上进行音频特征提取，最后使用卷积神经网络，SVM 等方法进行识别工作。实验主要从以下三个方面展开研究：

(1)中国传统乐器数据库的标注和整理

现有的音频数据库包含 78 件中国传统乐器，每个乐器文件夹大致包含 5 种类型的音频文件：基本演奏（弹，拉，吹，敲），音阶（不同速度，不同力度），演奏技巧，经典乐曲片段，乐器相关知识介绍等，共计 1630 个文件。每个音频的开头均包含有关该音频内容的描述，研究的主要工作就是将语音文本记录下来，对音频文件进行内容的标注。除此之外还存在一部分音频文件内包含多个内容，需要手动对其进行切割，然后进行标注工作完成数据库的整理。本部分研究的主要任务是对原始音频数据库进行去噪，分帧，切割和标注。主要指利用人工方式对每一个音频进行标注，然后利用端点检测以及人工检验的方式进行乐器音频的切割。

(2)中国传统乐器音频的乐器种类识别

中国传统乐器音频的乐器种类识别任务主要包含三个步骤。首先是音频特征的提取，本文主要提取了三种类型的特征：频谱质心和频谱延展度，频谱对比度，梅尔频谱。然后是神经网络的设计，这里主要参考了 VGGNet 的一些特点来设计卷积神经网络。最后通过不断对参数进行调整来实现乐器种类的识别。通过对使用三种特征后的乐器种类识别效果进行对比，发现采用梅尔频谱特征的识别准确率及识别效果最佳。同时实验基于 VGGNet 设计了不同层数的卷积神经网络，最终发现 8 层的卷积神经网络的分类效果最好。最后在 1500 个 batch 下，使用梅尔频谱特征，通过 8 层卷积神经网络进行识别，测试集准确率达到了 99.3%。

(3)中国传统乐器音频的演奏技巧识别

中国传统乐器音频的演奏技巧识别是在中国传统乐器音频的乐器种类识别的基础上进行研究的，主要分为以下两个方面：

① 单一乐器的演奏技巧识别

针对单一乐器的演奏技巧识别，前提是已经知道了该乐器的种类，然后就在该乐器范围内进行演奏技巧的识别。在进行本部分的识别工作时，首先要提取音频的梅尔频谱特征，然后利用预训练的神经网络 ResNet18 提取梅尔频谱图的特征，接着利用 SVM 对提取到的特征进行分类识别，最终识别准确率可以达到 99%以上。然后利用 T-SNE 对神经网络提取到的 512 维的特征进行降维，在二维平面观察特征提取的效果，同时利用 SVM 对降维后的特征进行分类识别，平均准确率可以达到 83.4%。

② 同一类乐器的演奏技巧识别

在进行单一乐器的演奏技巧识别时，如果数据库中没有待检测音频的演奏技巧，那么该演奏技巧一定无法识别。比如要识别曲笛的单吐技巧，但是数据库中却没有曲笛的单吐技巧的音频。此时如果对该音频进行识别，首先识别出乐器种类是曲笛，然后要识别其演奏技巧的话，只能给出一个近似的结果（可能是双吐）。因此，我们要进行同一类乐器的演奏技巧识别。这里的同一类是指我们根据乐器的演奏方式不同来划分的，共划分为了四大类：吹奏类，敲击类，拉弦类，弹拨类。曲笛属于吹奏类。在识别曲笛的演奏技巧时，可以用吹奏类乐器如梆笛的单吐技法音频来识别曲笛的单吐技法。这样我们的模型就具有更强的泛化性和鲁棒性。

具体实验方法是：先提取同一类乐器音频的梅尔频谱特征，然后搭建 8 层卷积神经网络进行训练，最后利用训练好的模型进行演奏技巧的识别。最终识别准确率为吹奏类乐器 95.7%，弹拨类乐器 82.2%，拉弦类乐器 88.3%，敲击类乐器 97.5%。

1.4 论文组织架构

本论文主要分为五个部分进行论述，具体如下：

第一章，引言。主要阐述论文的研究背景，国内外有关音频识别的研究现状和国内对于中国传统乐器音频识别的研究进展，还有研究的主要内容以及论文的整体架构。

第二章，乐器音频识别的理论基础。首先简单介绍了音频信号的基础知识。然后从整体分析了乐器音频识别的整个流程。接着介绍了论文中所涉及到的中国传统乐器音频数据库，然后解释一些论文中用到的乐器音频相关的时域及频域上的特征，常见的特征降维方法，最后说明了论文中用到的音频分类的方法。

第三章，基于卷积神经网络的中国传统乐器音频种类识别。首先介绍了音频信号的预处理及特征提取方法，然后具体描述了乐器音频种类识别算法的实现过程，并给出了最优的识别模型及识别结果，最后对识别结果从三个角度进行了分析。

第四章，基于卷积神经网络的中国传统乐器音频演奏技巧识别。首先介绍了单一乐器的演奏技巧识别过程，然后介绍了同一类乐器的演奏技巧识别方法。最后对两种识别方式进行了对比，同时给出了最后的识别结果分析。

第五章，对本研究的工作进行了总结，对未来在此基础上进行的实验进行了展望。

第二章　中国传统乐器音频识别的理论基础

本章主要介绍了论文的理论基础，主要分为两大部分。第一部分是音频信号的基础知识。第二部分是乐器音频识别的基础知识。第二部分首先总览了音频识别的整个架构流程，然后从数据库，特征提取，特征降维，分类识别四个方面进行了阐述。

2.1　音频信号的基础知识

对于一个音频信号来说，无论是音乐，还是人声，我们都会有不同的感受。我们也许会下意识的认为不同的声音，其本质构成是不同的。其实声音都是由振动产生的。每个时刻的振动都会有一个幅值，所以一个音频信号是由一个一维序列构成的。

对于一个声音，有四个比较常用的参数：时长，振幅，音高和音色。时长，简单来说就是一段声音的时间长度，也就是声音的持续时间。振幅，就是声音信号的强度，也就是声音振动的一种描述。音高，指声音的频率，音高越高，我们所感知的频率就越高。音色，主要是区别不同的声音，与声音的频谱有关，接下来我们将会提到。

对于一个音频信号，我们主要关心它的时长，采样率。时长和采样率决定了音频信号这个一维序列由多少个点组成。对于采样，在模拟信号转换为数字信号时，连续的信号同时也被转换为了离散的信号，采样就是每隔一定时间，从原音频序列中取出一个特定的点的过程。而采样率就是指在 1 秒的时间内，采集了多少个点。对于采样，必须要满足采样定理，才能保证不失真。如图 2-1 所示，我们在不同的范围下观察同一个曲笛的音频序列，我们可以在 2-1③发现每个具体的采样点。通过对比 2-1①和④我们可以发现音频序列点数和它的时长的关系（曲笛音频序列的采样率为 44100）。

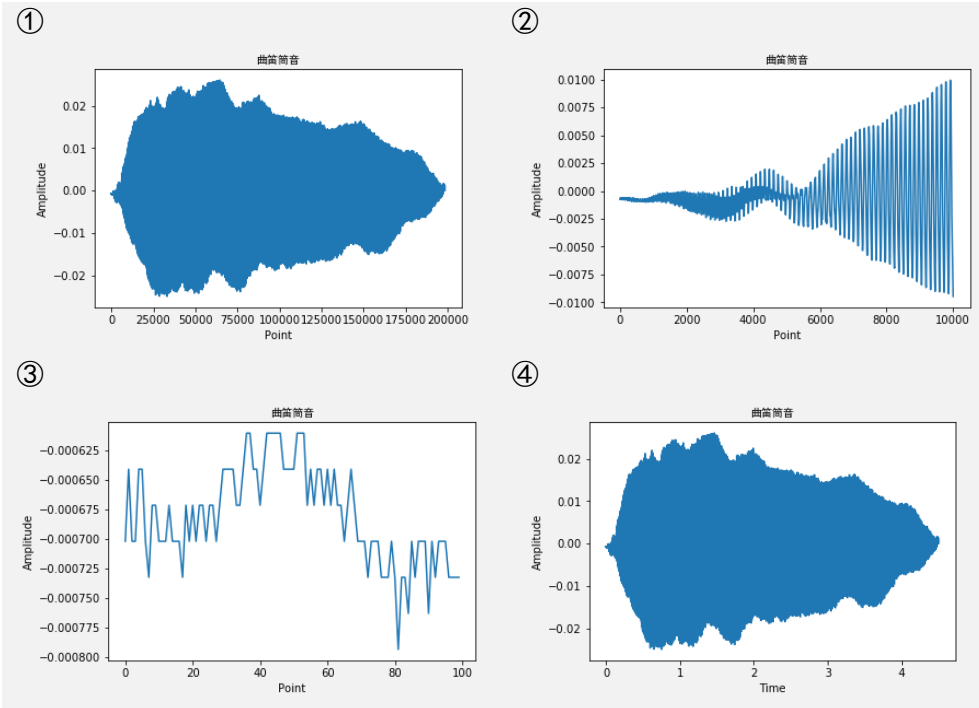

图 2-1 曲笛简音的音频序列

（图① 序列点；图② 序列点 0-10000；图③ 序列点 0-100；图④ 时间点）

对于一个音频序列，我们主要从时域和频域两个方面进行分析。有关时域我们其实在上文中已经进行了分析，而对于频域，我们这里用一个正弦函数的例子进行分析。

通过图 2-2①我们可以观察出函数$y = sinx + 0.3sin3x + 0.2sin5x$的每个子函数图像，这些子函数叠加形成了一个新的函数，如图 2-2②所示为新函数的图像。而时域和频域就是从两个不同的角度来观察这个新的函数。从图 2-3，我们可以观察出时域图与频域图的联系。由傅里叶变换可知，有一些函数可以通过傅里叶变换表示成三角函数的线性组合。通过频域，我们可以了解到一个音频序列的频率组成，这有助于分析它的音色。通过时域，我们可以了解到一个音频序列随时间的变化情况，这有助于从整体把握音频的走势。

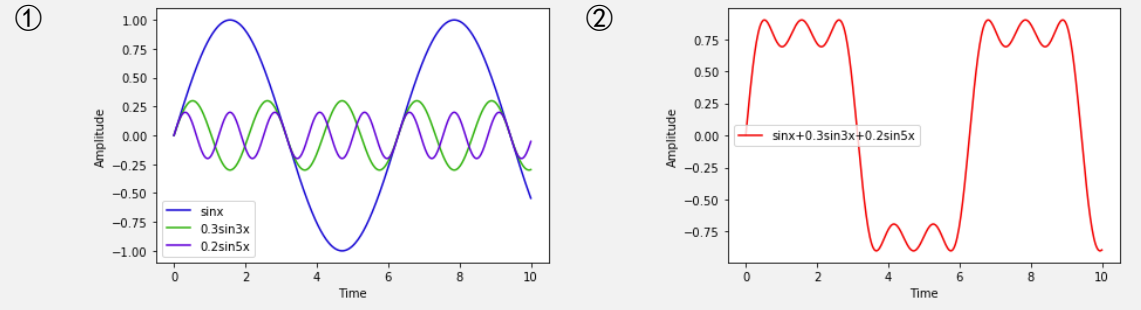

图 2-2 不同正弦函数的图像

（图① 子函数的图像；图② 和函数的图像）

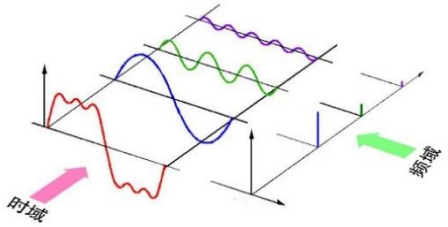

图 2-3 正弦函数时域及频域图像

2.2 乐器音频识别的系统架构

通过上文的描述，我们已经明白了什么是一个音频序列。接下来我们主要描述一下如何基于一个音频序列来实现识别效果。

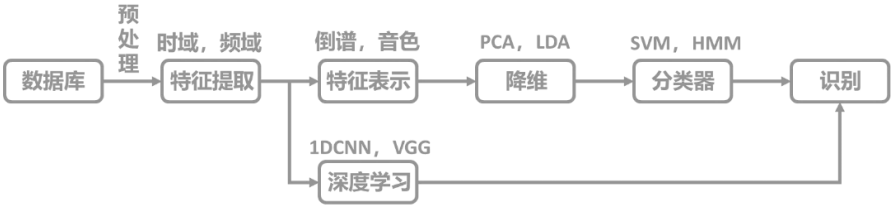

图 2-4 音频识别的一般流程

我们使用的主体方法是机器学习，机器学习是近年来较为热门的一个学科，它主要研究怎样让计算机来模拟人类的学习活动。它的核心是通过我们设定的规则，不断通过输入数据调整自身，最后达到模拟输出的目的。一个训练好的模型，我们只需给它输入需要识别的数据，它便会很快返回识别结果。我们的乐器音频识别主要是基于机器学习来实现的。它的主要流程如图 2-4 所示。

首先，进行机器学习需要大量的数据，只有大量的数据才能更好地帮助模型调节自身的参数。这里我们自己构建了一个用于训练和检验的数据库，具体的中国传统乐器音

频数据库在下文中将会详细介绍。有了大量的数据后，还无法直接用于训练。因为我们的训练数据都是统一的，要按照一定的规则进行输入，所以，我们要先对数据库中的数据进行预处理工作。对于音频序列来说，这不仅有一个常规的去噪分帧的过程，还有一个重要的添加标签的过程。

有了规整的数据之后，我们需要对数据进行特征提取。特征就是数据的某个方面最突出，最具代表性的特点。特征提取的目的主要因为原数据繁杂，冗余信息过多，机器学习无法直观学习到其本质，这其实与人的思维理解能力类似。关于音频序列的特征提取，我们主要从时域和频域两个方面出发。时域特征有 RMS 和短时过零率等，频域特征有频谱质心，频谱延展度，频谱对比度，梅尔频谱等。提取到原始音频的特征之后，我们会使用提取到的特征来表示原序列。特征表示的过程就是我们人为设定一些规则，加入我们自己的一些理解，对特征进行相应的处理来描述音频的主观特点。这里主要有倒谱表示和音色描述符等。

由于提取到的特征的维数可能过多，使得机器学习的运算量过于庞大，所以我们要进行降维操作。降维的主要目的是减少特征中的冗余信息引起的误差，降低运算复杂度。降维主要是降低特征的维数，其本质是用另一组特征向量表示新特征，相当于为特征提取特征。我们常用的降维方法为 PCA 或者 T-SNE。

在获得特征之后，我们就可以进行识别工作了。机器学习的识别算法众多，比较有代表性的就是支持向量机（SVM）。本文在演奏技巧识别中也会用到 SVM。随着深度学习的不断发展，我们也会使用另一种方法来进行音频识别工作，那就是直接将提取到的特征送入神经网络中进行训练，然后利用训练好的模型进行识别工作。对于音频序列，我们一般将其频谱特征送入卷积神经网络中进行训练获得模型。

接下来，我会从数据库，特征提取，降维，分类四个方面进行阐述，具体描述论文中的细节部分。

2.3 中国传统乐器音频数据库

2.3.1 数据库简介

原始音频数据库来源于中国音乐学院[14]，拥有 78 件中国传统乐器，1630 个音频文件。分别有吹奏类乐器 23 件，拉弦类乐器 13 件，弹拨类乐器 13 件，敲击类乐器 29 件。每个乐器文件夹大致包含 5 种类型的音频文件：基本演奏（弹，拉，吹，敲），音阶（不同速度，不同力度），演奏技巧，经典乐曲片段，乐器相关知识介绍等。每个音频的开头均有有关该音频内容的描述。具体原始音频数据库如图 2-5 所示

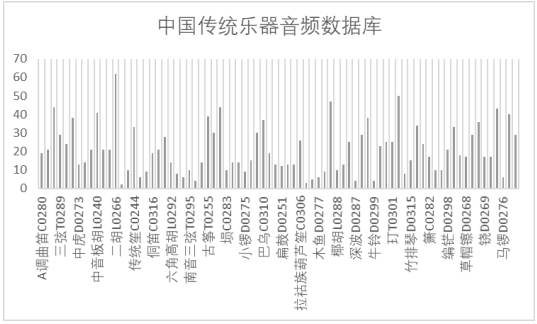

图 2-5 中国传统乐器音频原始数据库

2.3.2 数据库的标注与整理

我们的主要工作就是对原始数据库进行标注与整理，数据库的标注主要是将音频开头的语音描述进行记录，数据库的整理主要是将音频进行切割，切除空白部分和人声部分。具体做法是用人工方式对每一个音频进行标注，用端点检测对每一个音频进行切割，然后再用人工方式对端点检测的结果进行校准。端点检测的主要过程为：提取音频序列的 RMS（Root Mean Square）特征，然后进行滑动窗口，最后根据设定的判决条件进行端点的检测。端点检测的具体结果如图 2-6 所示。

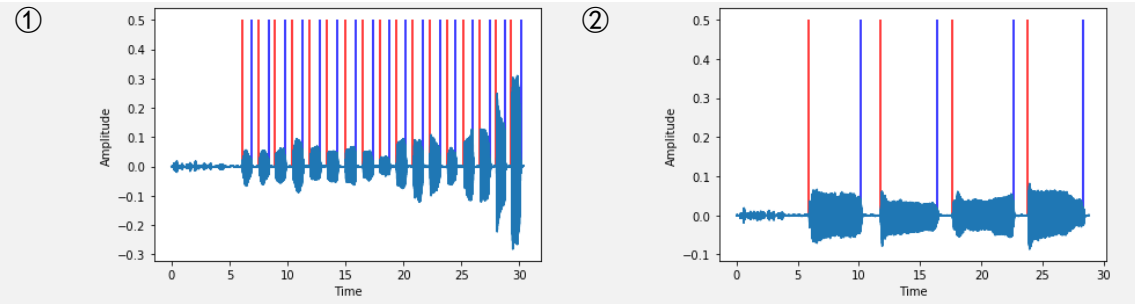

图 2-6 端点检测结果展示

(图① 端点检测结果 1；图② 端点检测结果 2)

经过整理，原始音频数据库共分为 1738 个音频，18819 个音频片段，乐器音频数量的均值为 240。整理后的新数据库如图 2-7 所示。整理后的音频数据库的信息使用一个 Excel 表格进行存储，如图 2-8 所示，分别保存音频的类别，路径，具体内容，是否人工核验，音频片段数量，音频片段切割点等信息。在使用中国传统乐器音频数据库时，我们只需要按照表格中的路径读入每个音频，然后根据其采样率和切割点就可以得到需要识别的音频片段了。

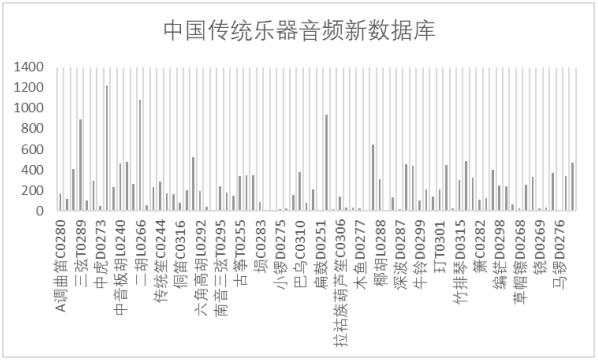

图 2-7 整理之后的中国传统乐器音频数据库

	case_id	category	route	content	check	number	cutpoint
0	1	A调曲笛C0280	乐器音频/A调曲笛C0280/C0280A2-1b1-2.wav	简音，自然衰减，力度弱	1.0	1	[[6.06, 10.33]]
1	2	A调曲笛C0280	乐器音频/A调曲笛C0280/C0280A2-1b2-2.wav	简音，自然衰减，力度强	1.0	1	[[5.2, 10.75]]
2	3	A调曲笛C0280	乐器音频/A调曲笛C0280/C0280A2-1b3-1.wav	简音，自然衰减，力度中强	1.0	1	[[6.41, 11.49]]
3	4	A调曲笛C0280	乐器音频/A调曲笛C0280/C0280A2-1c1-1.wav	音阶，慢速，力度弱	1.0	17	[[5.97, 6.71], [7.43, 8.06], [8.75, 9.33], [10...
4	5	A调曲笛C0280	乐器音频/A调曲笛C0280/C0280A2-1c2-1.wav	音阶，慢速，力度强	1.0	17	[[6.06, 6.87], [7.55, 8.38], [8.94, 9.82], [10...

图 2-8 中国传统乐器音频数据库具体记录（部分）

2.4 乐器音频的特征提取

上文中提到，对于一个音频序列，我们主要从时域及频域两个方面对其进行分析，本节，我会主要介绍一些常见的时频域的音频特征。

2.4.1 时域特征

(1) ADSR 介绍

对于一个音频序列，我们直观从时域分析，其最显著的特征就是振幅包络。振幅包络主要用来描述一段时间内音频信号的变化情况。对于乐器音频来说，其振幅包络也就是 ADSR 包络是其最重要的时域特征。ADSR 分别是指 Attack 激励，Decay 衰减，Sustain 持续，Release 释放四个过程。如图 2-9 所示为某一声音的 ADSR 包络。利用 ADSR 包络我们可以知道乐器音频时域的演奏状态。同时通过分析和使用 ADSR 包络我们自己也可以合成乐器的音色，我们称之为波表合成。对于图 2-9 的 ADSR 包络，0～1 表示“A”，1～2 表示“D”，2～5 表示“S”，5～7 表示“R”。我们经常使用 RMS 来提取 ADSR 包络。

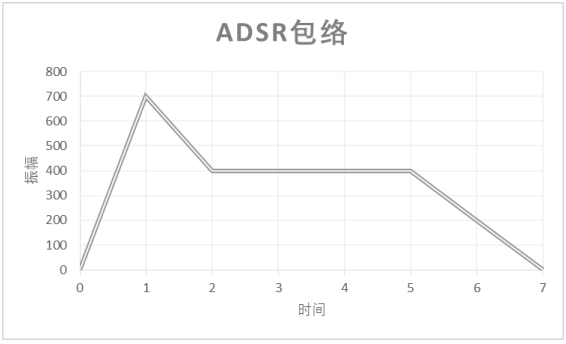

图 2-9 ADSR 包络示例

(2) RMS 均方根植

所谓 RMS 均方根植，其实就是对音频序列的某一部分的振幅进行运算。RMS 的具体计算如式 2-1 所示。其中 w 表示音频序列选择的窗口 window， i 表示每个被选择的音频采样点。

$$RMS = \sqrt{\frac{1}{w}\sum_{i=0}^{w-1} x^2[i]} \qquad \text{式（2-1）}$$

具体提取一个音频序列的 RMS 特征值时，还需要用到滑动窗口，对每个选择的音频序列的窗口进行 RMS 计算，最后同样会获得一个振幅包络。

(3) 短时过零率

顾名思义，短时过零率就是音频序列在一段时间内相邻采样点的值通过零值的频率，一般用来检测端点。短时过零率的计算如式 2-2，2-3 所示，其中 w 表示窗距，函数 y 表示原音频序列。

$$z(i) = 0.5 * \sum_{k=0}^{w} |sgn(y_i(k)) - sgn(y_i(k-1))| \qquad \text{式（2-2）}$$

$$sgn(x) = \begin{cases} 1, x \geq 0 \\ -1, x < 0 \end{cases} \qquad \text{式（2-3）}$$

2.4.2 频域特征

(1) 频谱质心与频谱延展度

频谱质心表示了频谱能量的集中位置，频谱延展度表示了在频谱质心周围的能量的分布状况。式 2-4 和 2-5 表示了每一帧频谱质心与频谱延展度的计算过程，其中 *S* 表示归一化后的频谱图，*freq* 是 S 的频率数组，*p* 值为一参数，这里取 2。

$$Centroid[t] = \sum S[k,t] * freq[k] / \sum s[k,t] \quad \text{式（2-4）}$$

$$Bandwidth[t] = (\sum S[k,t] * (freq[k,t] - Centroid[t])^{p})^{1/p} \quad \text{式（2-5）}$$

(2) 频谱对比度

Dan-Ning Jiang 等人[15]提到使用频谱对比度特征在音乐类型分类中比 MFCC 效果更好。其中频谱对比度和 MFCC 特征的提取过程如图 2-10 和 2-11 所示。频谱对比度主要考虑了每个子带的频谱峰值和谷值及其差异。

图 2-10 频谱对比度特征的提取过程

图 2-11 MFCC 特征的提取过程

频谱对比度的关键在于构建滤波器，本文中将频域划分为 8 个子带，0Hz～100Hz，100Hz～200Hz，200Hz～400Hz，400Hz～800Hz，800Hz～1600Hz，1600Hz～3200Hz，3200Hz～6400Hz，6400Hz～12800Hz，滤波器根据子带划分进行滤波。在计算频谱峰值和谷值时，我们分别取最大值和最小值的领域内的平均值作为结果。

具体计算方式如下：

① 首先将音频序列分帧，然后进行 FFT 运算，然后根据我们划分的八个子带，第 k 个子带的 FFT 分量是$\{x_{k,1}, x_{k,2}, \dots, x_{k,w}\}$，按照降序排列$\{y_{k,1}, y_{k,2}, \dots, y_{k,w}\}$。频谱峰值，谷值及其差异计算方式如式 2-6，2-7，2-8 所示。*m* 是一个参数，范围为 0.02～0.2。

② 然后对得到的特征向量进行对数运算，接着进行 KL 变换，最后得到频谱对比度特征。

$$Peak[k] = \log\left(\frac{1}{mN}\sum_{i=1}^{mN} y_{k,i}\right) \quad \text{式（2-6）}$$

$$Valley[k] = \log\left(\frac{1}{mN}\sum_{i=1}^{mN} y_{k,N-i+1}\right) \quad \text{式（2-7）}$$

$$d[k] = Peak[k] - Valley[k] \quad \text{式（2-8）}$$

(3) 梅尔频谱

梅尔频谱特征的获取流程如图 2-11 所示，其关键在于梅尔滤波器的构建。而梅尔

滤波器是根据梅尔刻度构建的。人耳对频率的感知是呈非线性变化的，只有经过梅尔刻度的转换，才能变为线性关系。普通频率刻度与梅尔频率刻度的转换如式 2-9 所示。据此构建了梅尔滤波器，如图 2-13 所示。本文中的梅尔滤波器由 64 个三角滤波器组成，高频处滤波器分布稀疏，阈值较小，低频处滤波器分布密集，阈值较大，这正是因为普通刻度和梅尔刻度呈对数关系。最终得到的梅尔频谱和原始频谱如图 2-12 所示。

梅尔频谱的具体计算方式如下：

① 对原始音频序列进行 STFT（短时傅里叶变换）运算。

② 构建梅尔滤波器，由 64 个三角滤波器组成。

③ 将短时傅里叶变换结果进行转置，然后与梅尔滤波器进行矩阵积，最后取对数。

$$mel(freq) = 2595 * log_{10}(1 + \frac{freq}{700}) \quad \text{式（2-9）}$$

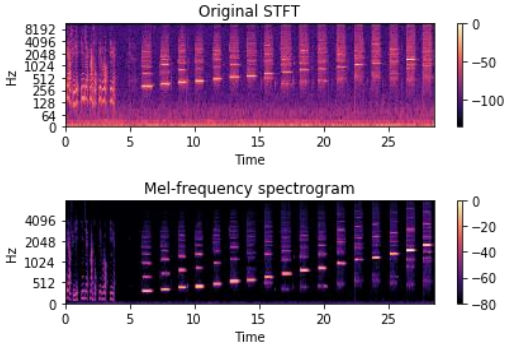

图 2-12 梅尔频谱与原始频谱的对比

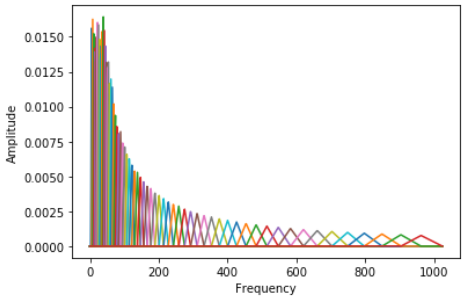

图 2-13 梅尔滤波器

2.5 乐器音频特征的降维方法

特征降维的方法较多，在本文中主要用到了 T-SNE。这主要是因为 T-SNE 降维效果比较好，研究比较新颖，然后可视化效果突出。

T-SNE 是由 L.J.P. van der Maaten 在 2008 年提出的[16]。T-SNE 的核心思想是在高维特征处使用高斯分布，在低维特征处使用 t 分布，使得距离可以用概率分布来描述。其中在高维特征处使用式（2-10）描述距离，在低维特征处使用式（2-11）描述距离。

$$P_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)}{\sum_{m!=n} \exp\left(-\|x_m - x_n\|^2 / 2\sigma^2\right)} \quad \text{式（2-10）}$$

$$Q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{m!=n} (1 + \|y_m - y_n\|^2)^{-1}} \quad \text{式（2-11）}$$

$$\frac{\nabla S}{\nabla y_i} = 4\sum_j (P_{ij} - Q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \quad \text{式（2-12）}$$

T-SNE 的算法流程如下：

(1) 设定输入向量 X，输出向量 Y

(2) 设置需要用到的参数：迭代数 N，学习率 L，变换动量 $a(t)$

(3) 初始化 Y。

(4) 迭代计算 Y 的值。具体计算方式为$y^t = y^{t-1} + L\frac{\nabla S}{\nabla y} + a(t)(y^{t-1} - y^{t-2})$

2.6 乐器音频分类的常用方法

2.6.1 支持向量机（SVM）分类

支持向量机（Support vector machines，SVM），是一种二分类器，主要思想是在两类特征数据中，选择一个超平面将两类特征数据最大限度地分隔开。其中的最大限度就突出了 SVM 具有很强的泛化性。式 2-13 就是我们所要求解的超平面。

$$y = w^T x + b = 0 \quad \text{式（2-13）}$$

SVM 的计算流程如下：

(1) 明确问题：$\max\gamma \quad s.t.\ \ y(w^T x + b) > \gamma$

(2) 将函数最小间隔缩小为 1：$\max\frac{1}{2}\|w\|^2 \quad s.t.\ \ y(w^T x + b) > 1$

(3) 构建拉格朗日函数：$L(w, b, a) = \frac{1}{2}\|w\|^2 - a(\ y(w^T x + b) - 1)$

(4) 求解拉格朗日函数，依次得到 a，w，b 的值。

由于 SVM 是一种二分类器，所以要解决多分类问题，有“一对一”和“一对多”两种思路。关于“一对一”，就是从多个类别中每次选出两个类进行分类，这样会构建 *n*(n-1)/2* 个分类器。关于“一对多”，就是固定一个类别，然后从剩余部分再选择一个类别，这样会构建 *n-1* 个分类器。虽然“一对多”运算量较小但是分类效果因所选类别而定，具有极大的偶然性，依赖于数据集本身的分布。而“一对一”的结果则会更为均匀，只是分类器数量是平方的级别，无法完成较多类别的分类。

2.6.2 神经网络分类

神经网络，深度学习的基础，是目前最为火热的机器学习算法。对于神经网络，我们可以把它想象成一个“黑盒”，我们只需要给它输入，它就会输出结果，但是一开始它的输出大概率不是我们想要的结果，这是因为它还没有进行“学习”。而神经网络的魅力也就在于它会自己进行“学习”。我们只需要对它的输出结果表明自己的看法，正确或错误，然后它就会不断调整自身的参数。随着这个过程的不断进行，它就会“学会”这项能力，然后我们就可以去使用它的这项能力。

在最近的研究中，Hendrik Purwins 等人[9]表明随着时代的不断发展神经网络已经取得三次卓越性的突破，分别是 1957 年感知器算法的提出，1986 年反向传播算法的提出和 2012 年语音识别和图像分类的成功。深度学习首先在图像处理中获得了成功，然后被广泛应用于音频信号等其他领域。

音频信号可以通过 STFT 从时域转换到频域得到二维频谱，但是其两个维度有特定的含义，分别表示时间和频率。在神经网络兴起之前，MFCC 一直是进行音频分析的关键特征，其主要流程是首先将音频转换到频域，然后进行对数运算，最后进行 DCT 变换。但是在神经网络中，DCT 变换会破坏频谱的空间信息，所以我们一般用省略掉 DCT 变换后的对数梅尔频谱作为神经网络音频分析的主要特征。

在本文中，我们主要使用卷积神经网络来进行识别工作。卷积神经网络相比于传统的全连接神经网络，大大降低了运算的复杂性，同时对于图像输入，我们不必经过繁琐的预处理工作，可以直接输入图像进行计算。卷积神经网络主要由卷积层和池化层组成。卷积层主要是通过卷积运算来提取特征，其中随着卷积层数的不断增加，相当于不断在特征的基础上提取特征，最后可以得到非常高层次的特征。池化层主要是通过下采样来对特征进行压缩处理。我们将不同的卷积层，池化层，全连接层和激活函数进行组装和搭配便可以搭建一个卷积神经网络。

第三章　基于卷积神经网络的中国传统乐器种类识别

本章主要解决中国传统乐器音频的乐器种类识别问题，主要基于卷积神经网络来实现的。本章首先描述了实验的预备工作，包括软硬件实验条件，乐器音频信号的预处理还有音频信号的特征提取。然后描述了整个基于卷积神经网络的乐器种类识别算法的主要流程，包括卷积神经网络的搭建，训练，调参，识别的过程。最后从三个角度对本实验的结果进行了分析。

3.1　软硬件实验条件

本论文的实验主要使用 Python 平台的 librosa 库来提取音频特征，主要使用 Pytorch 深度学习框架来进行仿真。具体的硬件条件如下表 3-1 所示。

表 3-1 硬件实验设备配置

名称	参数
CPU	i5-6300HQ 2.30GHz(4 核)
GPU	GeForce GTX 950M
RAM	16GB
HDD	1TB

3.2　乐器音频信号的预处理

本实验中预处理共分为两个步骤，主要是为了减少实验误差，提高实验的准确性。

(1)重采样

本实验取 44100Hz 为基准，对不同于该采样率的音频进行重采样。对于本音频数据库，所使用的采样均为上采样。上采样主要使用 *sinc* 函数进行内插[17]。

(2)归一化

归一化操作是将音频序列的幅值缩小到-1～1 之间，简化运算，加快求解速度。这里为了保持原序列的比例，直接对整个序列除以序列绝对值的最大值。

3.3　乐器音频信号的特征提取

对于乐器音频信号的特征提取，我们在上章中已经从时域及频域两个方面进行了讨论，因为我们主要使用卷积神经网络进行识别，所以我们主要使用频域的特征。在下文中我们主要使用了频谱质心与频谱延展度，频谱对比度，梅尔频谱三种特征。

3.3.1　频谱质心与频谱延展度

频谱质心表示了频谱的集中位置，而频谱延展度则表示了频谱质心的周围的能量分布状况。

我们的音频是 44100Hz 的采样率，所以我选择使用 2048 点的窗口和 1024 点的跳距，利用滑动窗口进行特征提取。如图 3-1 所示，我们对原始音频依照算法提取了频谱质心与频谱延展度特征，然后又加入了两种特征序列的一阶导数作为两种新特征表示序列的变换情况。最终我们用一个 4 维向量表示原序列。在图 3-1 的右半部分中，cent 和 band 分别表示频谱质心与频谱延展度，dcent 和 dband 分别表示频谱质心的一阶导和频谱延展度的一阶导。

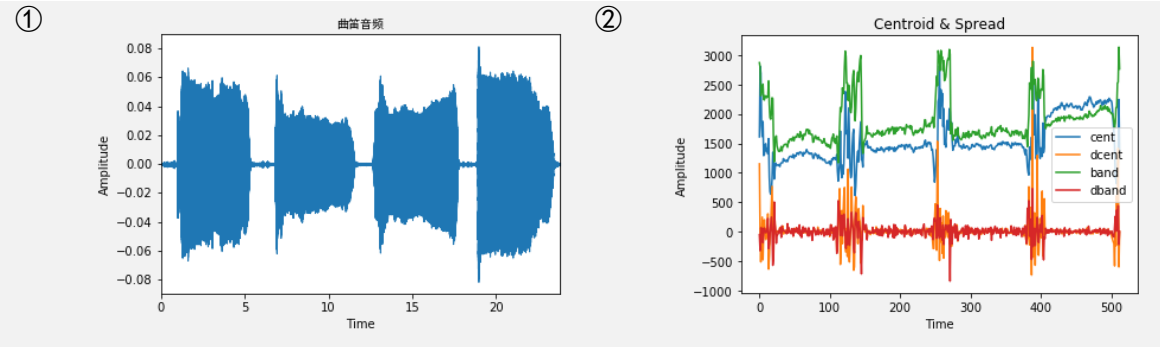

图 3-1 原始音频与其频谱质心与频谱延展度特征

（图① 曲笛原始音频；图② 提取到的特征）

3.3.2 频谱对比度

有关频谱对比度特征，我们使用了 8 个子频域，在 0，100，200，400，800，1600，3200，6400，12800 之间。

我们的音频是 44100Hz 的采样率，所以我选择使用 2048 点的窗口和 1024 点的跳距，利用滑动窗口进行特征提取。如图 3-2 所示，我们选取了曲笛的单吐技法的音频（前面没有规律的一小段是人声部分），对原始音频依照算法提取了频谱对比度特征。最终我们得到了一个 8 维的特征向量表示原音频序列。在图 3-2 的右半部分中，上半部分表示原始音频的频谱图，下半部分表示原始音频的频谱对比度特征。

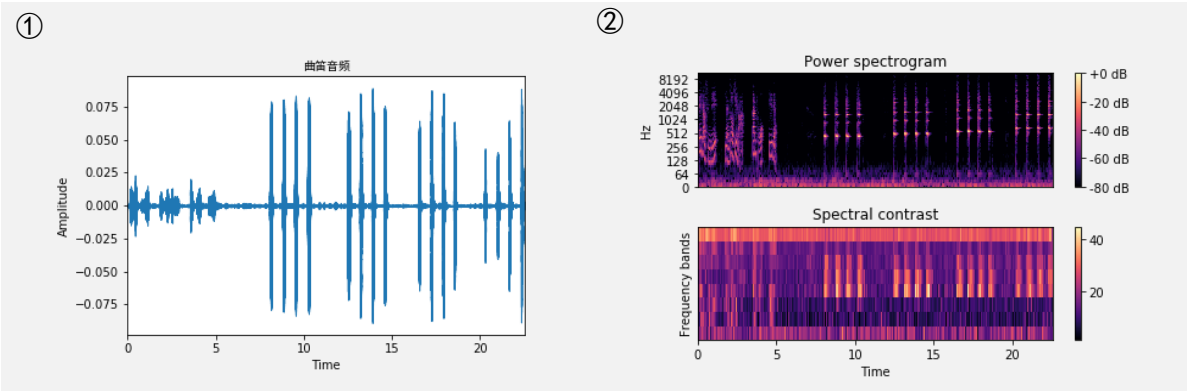

图 3-2 原始音频与其频谱对比度特征

（图① 曲笛原始音频；图② 提取到的特征）

3.3.3 梅尔频谱

对于梅尔频谱，我们选择了 64 个三角滤波器作为梅尔滤波器，最大频率为 22500，最小频率为 50，窗距为 2048 点，跳距为 1024 点。

我们的音频采样率是 44100Hz。我们首先使用汉明窗，窗距 2048 点，跳距 1024 点对原序列进行 STFT（短时傅里叶）变换，然后利用构建的梅尔滤波器过滤，最后进行对数运算。如图 3-3 所示，我们选取了曲笛音阶的音频（前面没有规律的一小段是人声部分），对原始音频依照算法提取了梅尔特征。最终我们得到了一个 64 维的特征向量表示原音频序列。在图 3-3 的右半部分中，上半部分表示原始音频经过 STFT 得到的频谱图，下半部分表示原始音频的梅尔频谱特征，我们可以从音阶观察出梅尔刻度与一般刻度的关系。

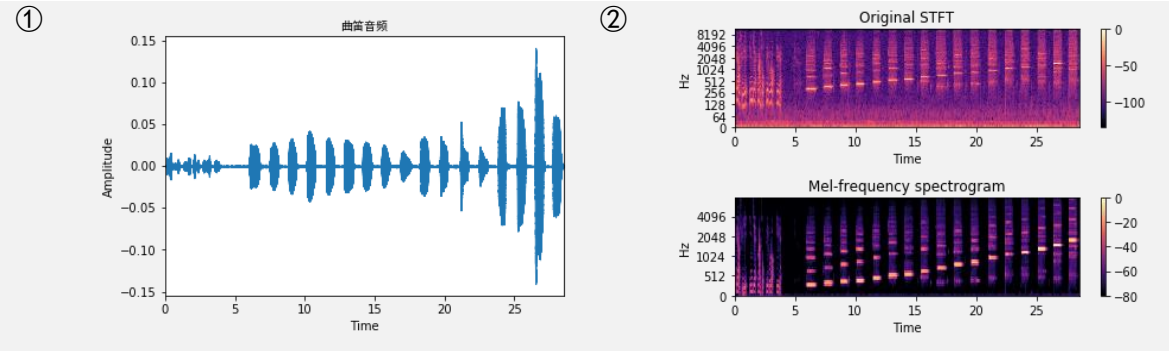

图 3-3 原始音频与其梅尔频谱特征

（图① 曲笛原始音频；图② 提取到的特征）

3.4 基于卷积神经网络的中国传统乐器种类识别算法

3.4.1 算法主要流程

中国传统乐器音频的乐器种类识别算法的主要流程如图 3-4 所示，首先我们将提取到的特征数据集划分为训练集和测试集，训练集用于调整模型参数，测试集用于评估模型质量。这里我们按照 8：2 的比例进行划分，训练集 15020 条，测试集 3900 条。然后我们分批将训练集中的特征送入卷积神经网络中进行训练，根据识别结果和正确结果的关系不断调整模型参数。在训练一定的轮数后，利用测试集检验模型的优劣性。这里我们选取梅尔频谱特征作为待使用的特征集。

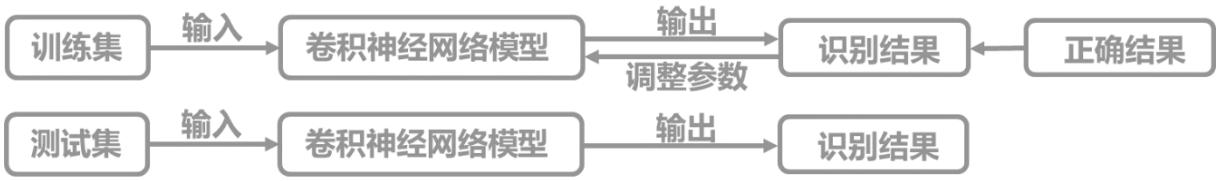

图 3-4 基于卷积神经网络的中国传统乐器种类识别算法流程

3.4.2 卷积神经网络的搭建

本文的卷积神经网络主要基于 VGGNet[18]搭建。VGG 网络的主要结构是由输入，卷积层，全连接层，输出连接组成。VGG 网络的特点如下：

(1) 使用较小的卷积核和池化核进行运算。其中主要使用 2*2 大小的池化核和 3*3 大小的卷积核。

(2) 使用几个较小的 3*3 卷积核来代替一个大的 5*5 或者 7*7 的卷积核。

我们基于 VGG 网络的特点，重新设计了八层的卷积神经网络。如表 3-3 所示，其中每一个卷积块（ConvBlock）都与表 3-2 卷积块 1 的结构类似。我们在每一个卷积块中使用两个较小的 3*3 的卷积核，每次进行卷积操作之后都进行 Batch Normal 操作。在经过最后一个卷积块之后，特征向量变为 512*8*4 维，这时我们使用一个 8*4 的池化核使特征向量集中到同一维度上，然后再经过全连接层，最后进行 SoftMax 回归。

表 3-2 卷积块 1 的结构

Input Size	net	size
[1, 128, 64]	Conv	[64, 1, 3, 3]
	Batch Normal	[64]
[64, 128, 64]	Conv	[64, 64, 3, 3]
	Batch Normal	[64]
[64, 128, 64]	Pool	kernel_size=(2, 2), stride=(2, 2)

表 3-3 八层的卷积神经网络结构

Input Size	net	size
[1, 128, 64]	ConvBlock1	
[64, 64, 32]	ConvBlock1	
[128, 32, 16]	ConvBlock1	
[256, 16, 8]	ConvBlock1	
[512, 8, 4]	Pool	kernel_size=(8,4)
[512, 1, 1]	View	
[512]	Linear	[512, 78]
[78]	SoftMax	
Parameters	4726158	

3.4.3 卷积神经网络的训练

在进行卷积神经网络的训练时，有以下几个步骤：

(1)设定优化器，这里选择 Adam 优化器。Adam 优化器可以自己调整学习率，而且实现较为简单，计算更为高效。

(2)设定损失函数，这里选择 NLLLoss（最大似然/log 似然损失函数）。具体损失函数的计算如式（3-1）所示。y_i表示第 i 个神经元输出值。t_i表示第 i 个神经元输出对应真实值的 one-hot 编码。

$$L = -\sum_i t_i log y_i \qquad \text{式（3-1）}$$

(3)调整输入特征，由于提取特征时音频时长的不固定，所以利用 STFT 等操作提取特征后，获得的特征为 t*64 维，其中 t 值与音频的时长有关。为了在神经网络中统一输入，所以我们要设定某一固定值，经过测试选取 128。将获得的特征均调整为 128*64 维。我的具体做法是对于所有特征每 128 个点进行切割，然后不足 128 个点的特征进行填充（填充方式为循环）。

(4)设定 Batch Size 为 64，即每次向卷积神经网络中投喂 64 组数据进行训练。

(5)每 200 个 Batch 进行一次测试集预测，每 1000 个 batch 保存一次模型。

经过 5000 个 Batch 的训练，具体的训练效果如图 3-5 和 3-6 所示。准确率达到了 99.6%，k 值（k 在这里取 3）平均准确率达到了 99.8%。在 1500 个 Batch 左右，损失值 Loss 下降为了 0.0004。

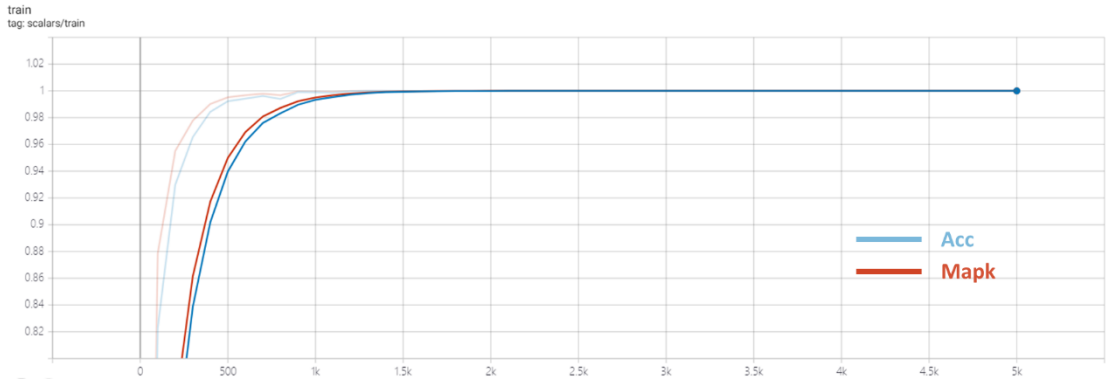

图 3-5 5000 个 Batch 下的准确率与 k 值平均准确率

图 3-6 5000 个 Batch 下的 Loss 值变化

3.4.4 卷积神经网络的识别结果

我们的测试集音频数量为 3799，最终有 25 个音频没有分类正确。如图 3-7 所示，通过混淆矩阵我们可以看出不同演奏方式乐器的大类之间几乎没有分类错误，即敲击类，拉弦类，吹奏类，弹拨类这四类乐器能基本进行正确识别。

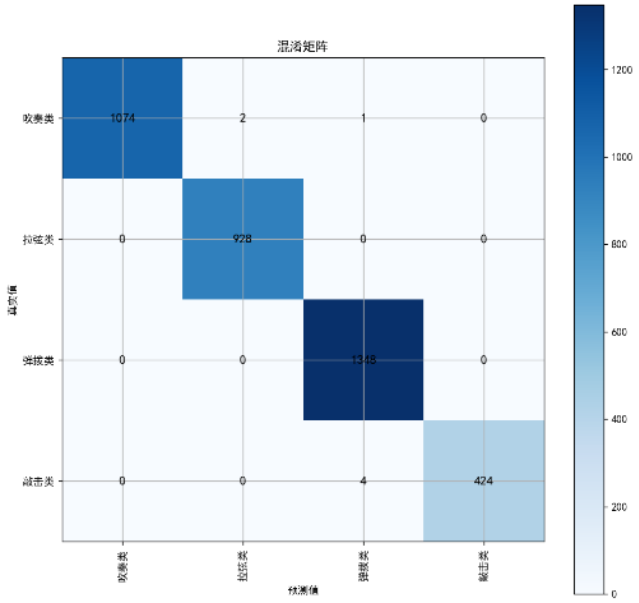

图 3-7 四类不同演奏方式乐器的混淆矩阵

每大类乐器之间的混淆矩阵如图 3-8 所示。吹奏类乐器内部有 4 个分类错误，敲击类乐器内部有 5 个分类错误，弹拨类乐器内部有 9 个分类错误，拉弦类乐器全部分类正确。

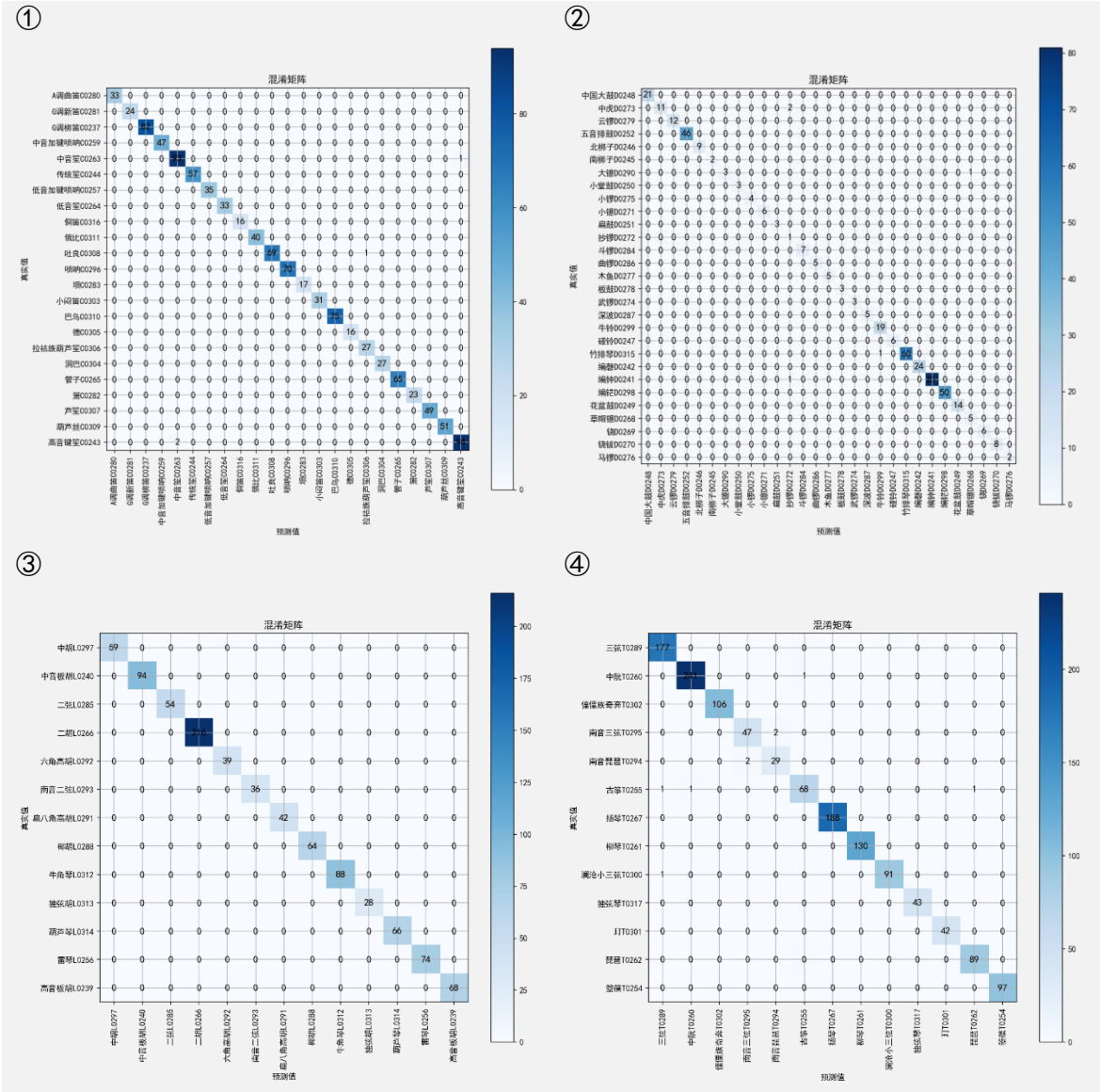

图 3-8 每种不同方式演奏乐器内部混淆矩阵

（图① 吹奏类；图② 敲击类；图③ 拉弦类；图④ 弹拨类）

3.5 结果分析

3.5.1 基于不同输入特征的评价

我们一共使用了三种不同的特征作为卷积神经网络的输入特征，分别是四维的频谱质心，频谱延展度，频谱质心的一阶导，频谱延展度的一阶导组合而成的特征；八维的频谱对比度特征；六十四维的梅尔频谱特征。其中在使用四维的特征时，需要对卷积神经网络做一些调整（改变池化的次数）。

通过表 3-4 我们可以看出，虽然三种特征在训练集上均可达到接近 100%的准确率，但是存在过拟合的状况。我们在测试集上进行实验，发现梅尔频谱的准确率最高，频谱对比度的准确率次之，频谱质心与频谱延展度特征的准确率最差。因此使用梅尔频谱特征作为输入特征，得到的卷积神经网络模型的效果最好。

表 3-4 不同输入特征的训练结果

	Loss	训练集准确率	测试集准确率
频谱质心与频谱延展度	0.01	99.9%	87.9%
频谱对比度	0.009	99.9%	94.4%
梅尔频谱	0.0004	99.9%	99.3%

3.5.2 基于不同层数卷积神经网络的评价

我们根据上述实验中 8 层卷积神经网络的搭建方法，又搭建了 4 层和 6 层的卷积神经网络。

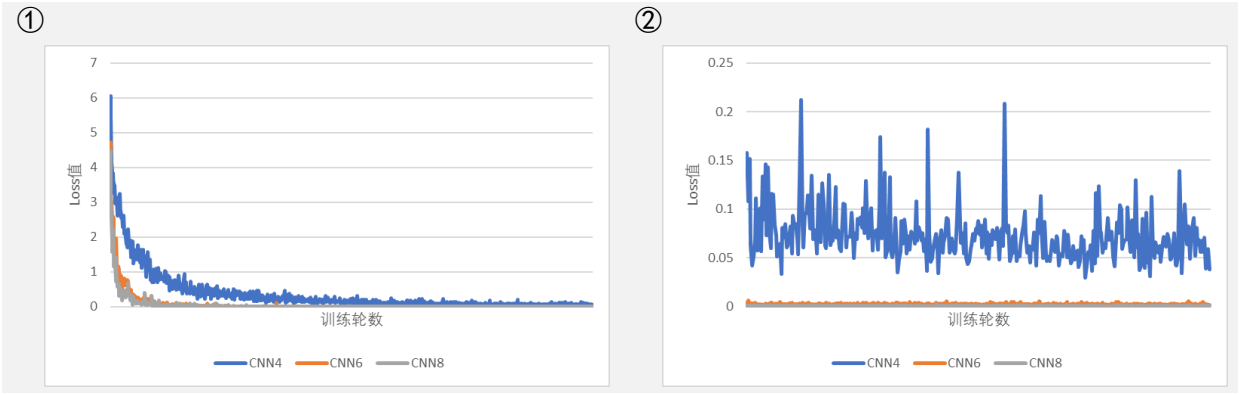

图 3-9 不同层数的卷积神经网络训练效果

（图① 0～3000 个 Batch；图② 2000～3000 个 Batch）

最终经过了 5000 个 Batch 的训练，训练集均可以达到 99.9%的准确率，而 Loss 值的变化不太相同。图 3-9 展示了 0～3000 个 Batch 中 Loss 值的变化情况。我们可以看出 CNN6 和 CNN8 的收敛速度较快，而 CNN4 的收敛速度较慢。同时 CNN6 和 CNN8 达到收敛时的 Loss 值要低于 CNN4。通过表 3-5 我们可以看出 CNN6 和 CNN8 的模型识别效果相差较小，而 CNN4 的模型识别效果与它们相比要差一些。

表 3-5 不同层数的卷积神经网络训练效果

	Loss	训练集准确率	测试集准确率
CNN4	0.04	99.9%	95.5%
CNN6	0.001	99.9%	98.9%
CNN8	0.0004	99.9%	99.3%

3.5.3 最终识别结果的评价

为了防止出现过拟合的问题，我们又使用 8 层的卷积神经网络进行了测试，训练集与测试集的划分方式为 1：1。其中训练集和测试集均为 9426 个。通过图 3-10 我们可以发现不同数据集划分下神经网络收敛速度相差不远，最终的 Loss 值也基本相同。当数据集划分为 8：2 时，测试集准确率为 99.3%，当数据集划分为 1：1 时，测试集准确率为 98.9%。因此可以判断出模型的过拟合状况较小，符合预期目的。

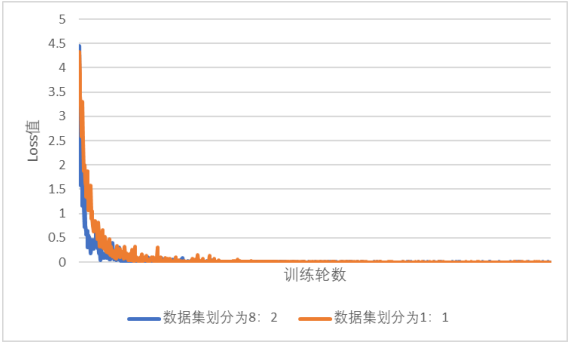

图 3-10 3000 个 Batch 下不同数据集划分的 Loss 值变化

在最终的识别效果中，有以下几个特点：

(1)同一乐器的不同部位发声均可正确识别。如二胡的弓杆击琴筒技法与二胡拉弦的声音完全不同，但是在本模型中仍然可以正确识别为二胡。还有琵琶的击板技法也可以得到正确的识别。

(2)频谱图类似的乐器均可正确识别。如扁鼓的鼓钉技法与扬琴的轮音技法的频谱图类似，但是仍然可以识别为扁鼓。

(3)声音类似的乐器均可正确识别。如萧和 G 调新笛，从人耳触觉来看几乎没有差别，但是模型仍然可以正确区分。

(4)弹拨类乐器与敲击类乐器容易分错。如草帽镲和大镲，从听觉来看，这两种乐器几乎没有差别，而且发声时间较短，ADSR 包络类似，音频数量较少，所以即使是神经网络也难以辨别。

3.6 小结

本章主要基于卷积神经网络进行中国传统乐器音频的乐器种类识别工作，最终使用 8：2 的数据集划分，利用梅尔频谱特征作为输入，训练 8 层的卷积神经网络作为最终模型，在测试集上取得了 99.3%的准确率。同时通过实际检验，发现模型识别效果符合预期目的。

第四章　基于卷积神经网络的中国传统乐器演奏技巧识别

对于中国传统乐器音频的演奏技巧识别，本章主要分为两类来进行实现。首先进行单一乐器的演奏技巧识别。然后考虑到模型的泛化性问题，为了让没有包含在数据库中的音频也可以得到准确的识别，所以接着进行了同一类乐器的演奏技巧识别。例如数据库中没有曲笛的单吐技巧，那我们使用单一乐器演奏技巧识别无法出曲笛的单吐技巧，而我们使用同一类乐器的演奏技巧识别可以通过新笛的单吐技巧识别出曲笛的单吐技巧，从而增强了模型的鲁棒性。

4.1　乐器音频信号的特征提取

在中国传统乐器音频的乐器种类识别中，梅尔频谱的表现要明显优于频谱对比度，频谱质心与频谱延展度。所以我们在中国传统乐器音频演奏技巧识别中，仍然选择使用梅尔频谱特征。

与上文的特征处理不同，上文中我们将梅尔频谱特征的时间维度每满 128 个点进行切割，因为在时间维度上，每一帧对应的音频种类是相同的。而进行本章的演奏技巧识别时，每一种演奏技巧在时间上是不同的（如果将某一连续的演奏技巧从时间上分割开，那么分割开的部分的频谱图是有较大差异的），同时同一种演奏技巧的时间长度也不一定相同（对于相同的演奏技巧，时长由人为控制）。所以在进行本章乐器音频的演奏技巧识别时，是将不同的梅尔频谱特征统一缩放到 224*224 大小。这里的梅尔频谱图缩放采用双线性插值。具体的双线性插值方式如式 4-1 所示，该式是将插值点移动到(0,0),(1,0),(0,1),(1,1)之间进行插值。

$$g(x,y) = f(0,0)(1-x)(1-y) + f(1,0)x(1-y) + f(0,1)y(1-x) + f(1,1)xy \quad \text{式（4-1）}$$

4.2　单一乐器的演奏技巧识别算法

4.2.1　算法主要流程

对于单一乐器音频的演奏技巧识别算法流程如图 4-1 所示。主要步骤如下：

(1) 提取某一种类乐器音频的梅尔频谱特征。

(2) 调整提取到的梅尔频谱特征。这里是将梅尔频谱图缩放到 224*224 的大小。

(3) 利用预训练的 ResNet 卷积神经网络提取梅尔频谱图的特征。这里的输出是一个 512 维的特征。

(4) 利用 SVM 进行特征的分类识别。

图 4-1 单一乐器音频演奏技巧识别算法流程

4.2.2　算法具体实现

梅尔频谱特征的提取在上文中已经进行了介绍，下面主要介绍利用 ResNet[19]卷积神经网络提取梅尔频谱图的特征。ResNet 网络又名残差网络，主要核心是残差模块。具体的残差模块类似图 4-2 所示，其中共 4 个残差模块，主要是为了解决梯度消失的问题。

我们将一个预训练的 18 层 ResNet 网络模型移除最后的全连接层，直接输出最后的 512 维特征。表 4-1 是我们获得的预训练的 17 层卷积神经网络模型，图 4-2 是 17 层卷积神经网络中的前两个卷积块。这里我们为了保持输入一致，需要将梅尔频谱图调整为 3*224*224.

使用预训练 ResNet 网络来提取频谱图的特征实际是一种迁移学习。对于每一单个乐器的演奏技巧，其数据量过小，无法训练出效果优秀的神经网络模型（神经网络欠拟合）。所以我们需要从之前的梅尔频谱图特征基础上再进行特征提取，然后进行机器学习。由于 ResNet 之前在 ImageNet 上表现突出，所以使用预训练 ResNet18 模型提取图片（我们将梅尔频谱特征调整为一张固定大小的图片）的特征，在移除最后的全连接层后，直接返回 512 维向量，然后我们就可以使用 512 维特征进行机器学习了。

表 4-1 十七层卷积神经网络结构

Input Size	net	size
[3, 224, 224]	Conv	7*7,64,stride=2
[64, 112, 112]	Max Pool	3*3,stride=2
[64, 56, 56]	ConvBlock1	
[64, 56, 56]	ConvBlock2	
[128, 28, 28]	ConvBlock3	
[256, 14, 14]	ConvBlock4	
[512, 7, 7]	Pool	7*7
[512, 1, 1]		

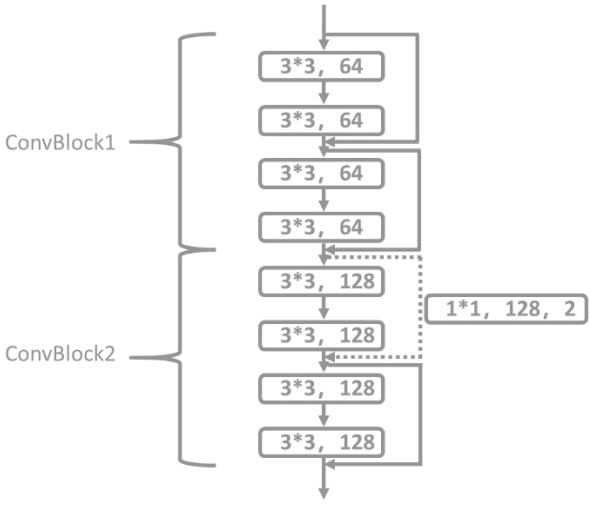

图 4-2 17 层卷积神经网络中的部分卷积块

利用神经网络模型提取完特征之后，我们使用 SVM 进行识别。这里惩罚参数设为 1.0，核函数选为 rbf（径向基函数），核函数参数设为 0.1。

4.2.3 单一乐器的演奏技巧识别结果及结果分析

我们直接对 512 维特征进行了 SVM 识别，识别准确率均在 99%以上。这表明我们对于单一乐器的识别准确率在 99%以上。

随后我们将提取到的 512 维特征，利用 T-SNE 降维到 2 维，然后再利用 SVM 进行识别。具体的识别准确率和对应音频数量如图 4-3 所示。经过分析我们可以发现以下两个特点：

(1) 敲击类乐器演奏技巧识别准确率低，这主要是因为对于敲击类乐器，不同的演奏技巧相差较小。中虎准确率为 0.5，大镲准确率为 0.428，小堂鼓准确率为 0.428，小镲准确率为 0.28。

(2) 音频数量与识别准确率存在一定关系。音频数量较少的话，音频准确率都较低。一般敲击类乐器音频数量较少，所以我们可以发现它们的识别准确率较低。

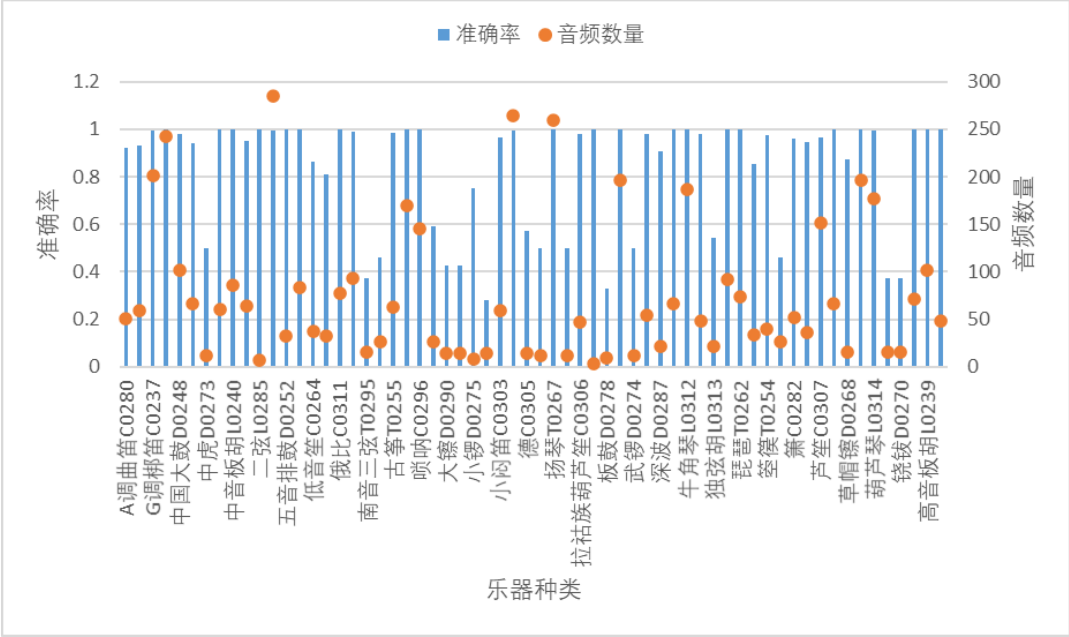

图 4-3 特征降维后的识别准确率及音频数量

同时通过图 4-4 的四种乐器，我们可以发现降维之后的演奏技巧特征还是聚集分布的。这不仅能推测同一乐器不同演奏技巧之间的相似度，而且极大地证明了利用 ResNet 网络提取频谱特征的有效性与可靠性。例如图 4-4①，该图表示乐器吐良的演奏技巧特征，具体对应如下：'筒音'，1；'演奏技法颤音'，2；'演奏技法单吐'，3；'演奏技法双吐'，4；'演奏技法三吐'，5；'演奏技法气震音'，6；'演奏技法气滑音'，7。通过图中点的分布，我们可以观察出每种特征之间的关系。

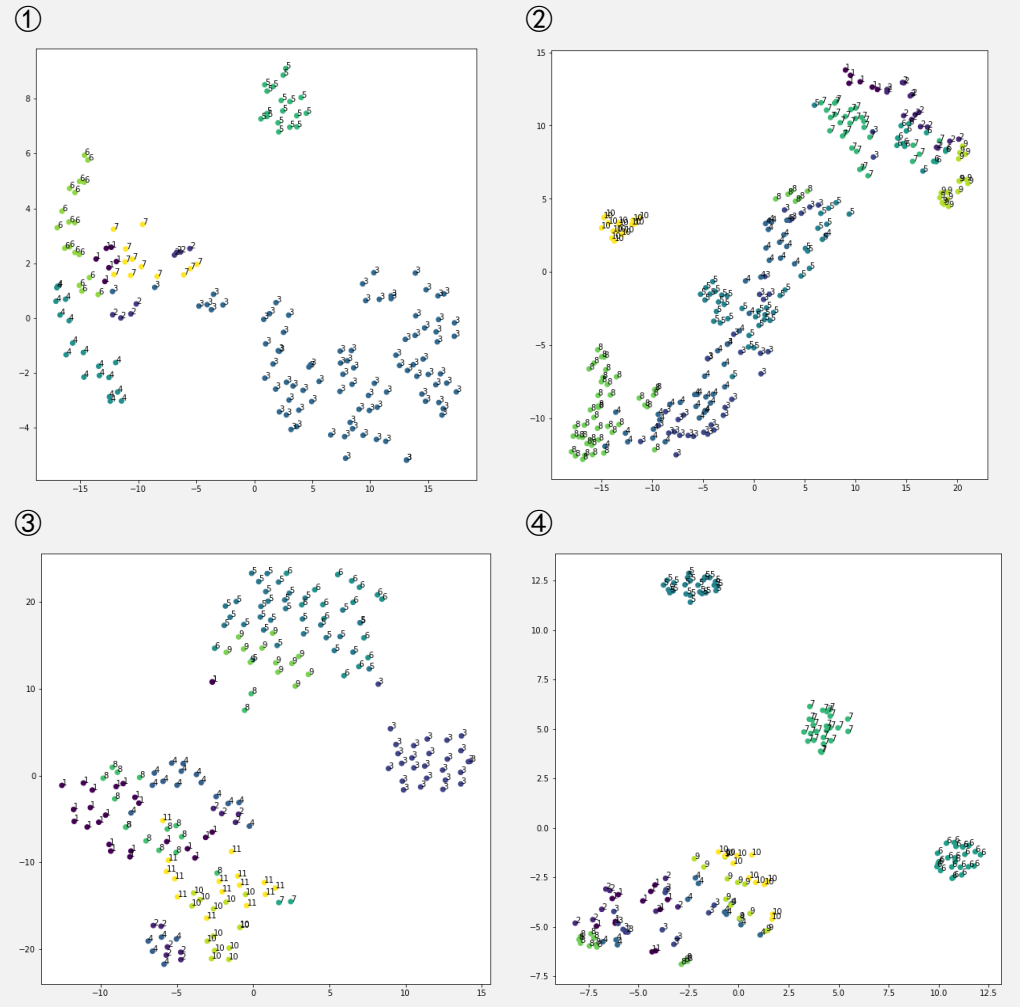

图 4-4 不同乐器降维后的特征分布

（图① 吐良；图② 巴乌；图③ 牛角琴；图④ 芦笙）

4.3 同一类乐器的演奏技巧识别算法

4.3.1 数据集介绍

对于同一类乐器的演奏技巧识别，我们将中国传统乐器音频数据库按照演奏技法划分为了四大类：吹奏类，弹拨类，拉弦类，敲击类。同时在每类音频数据库中只保留了基本技法（筒音，弹，空弦音，敲）和演奏技法。吹奏类共 57 种演奏技巧，弹拨类共 69 种演奏技巧，拉弦类共 64 种演奏技巧，敲击类共 35 种演奏技巧。

4.3.2 算法主要流程

对于同一类乐器的演奏技巧识别，主要有两种算法：

(1) 和单一乐器的演奏技巧识别算法类似，首先提取同一类乐器音频的梅尔频谱特征，然后利用预训练的 ResNet 网络提取其梅尔频谱图的特征，最后使用 SVM 进行识别。通过实验发现四大类乐器的识别准确率如下：吹奏类 86.8%，弹拨类 66.5%，拉弦类 75.2%，敲击类 80.6%。这个效果并不是很理想，所以我们使用下面的方法进行同一类乐器音频的演奏技巧识别。

(2) 首先提取同一类乐器音频的梅尔频谱特征，然后自己搭建卷积神经网络，最后训练卷积神经网络模型进行识别。

4.3.3 卷积神经网络的训练

卷积神经网络仍然使用中国传统乐器音频的乐器种类识别中的 8 层卷积神经网络。然后将数据集按照 8：2 划分为训练集和测试集。我们一共训练了 100 个 epoch。其中四类乐器均得到收敛，敲击类和吹奏类在收敛过程中出现了大幅波动，最后四类乐器的训练集准确率均接近 100%。如图 4-5，4-6 分别是四大类乐器的 Loss 值和训练集准确率的变化情况。

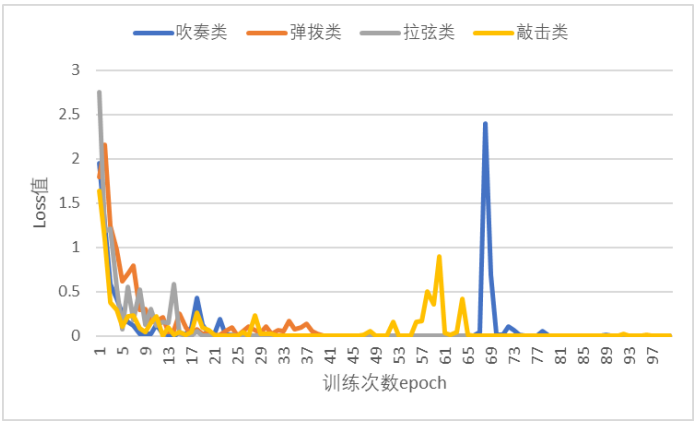

图 4-5 四大类乐器的 Loss 值变化

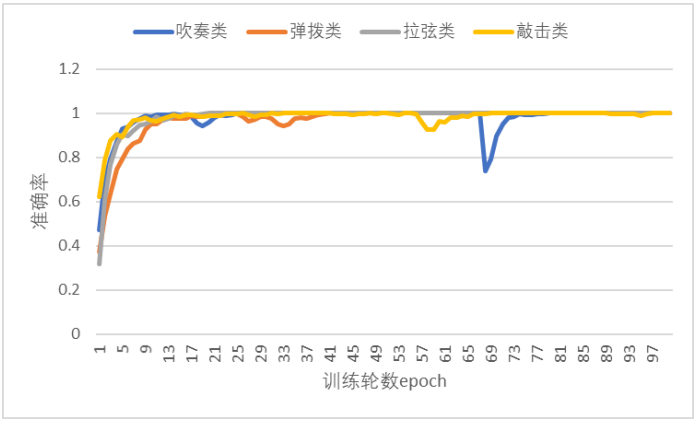

图 4-6 四大类乐器的训练集准确率变化

4.3.4 同一类乐器的演奏技巧识别结果

我们对进行 100 个 epoch 训练时，四类乐器测试集准确率的变化情况进行了研究，具体如下图 4-7 所示。最终各大类乐器音频演奏技巧识别准确率趋于稳定，其中吹奏类乐器 95.7%，弹拨类乐器 82.2%，拉弦类乐器 88.3%，敲击类乐器 97.5%。

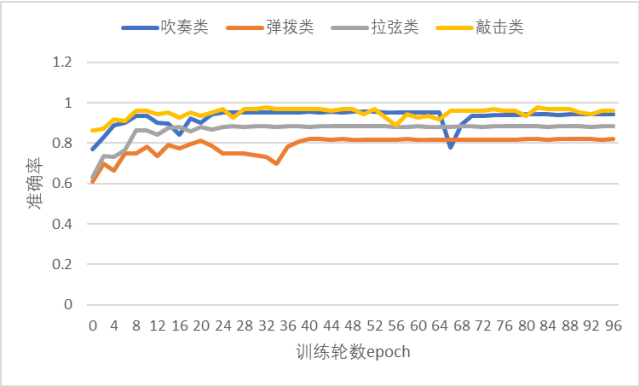

图 4-7 四大类乐器的测试集准确率变化

4.4 小结

我们针对中国传统乐器音频的演奏技巧识别问题，提出了单一乐器的演奏技巧识别和同一类乐器的演奏技巧识别两种方案。其中单一乐器的演奏技巧识别可以达到 99%准确率，同一类乐器的演奏技巧识别可以达到 90.9%的平均准确率。同一类乐器的演奏技巧识别可以增强演奏技巧识别的泛化性，调高识别的准确性。

第五章　总结与展望

5.1　总结

随着深度学习技术的不断发展，人们逐渐将深度学习应用于各种领域。本文主要基于卷积神经网络研究了中国传统乐器的音频识别工作。主要工作总结有三个方面：

首先，国外对于音频识别的研究已有多年，不仅有成熟的音频特征提取算法，而且也有可供参考的乐器自动分类流程。国内对于音频识别的研究也有不少进展，但是针对于中国传统乐器的研究较少，而且实验中所使用的数据库乐器种类较少，音频片段较少，无法进行大规模的中国传统乐器识别工作。因此我们为了填补这一片空白，自己构建了中国传统乐器音频数据库并给出了基于该数据库的音频识别的研究。我们构建的中国传统乐器音频数据库包含 78 件中国传统乐器，平均每种乐器有 200 个音频片段，同时我们为每一个音频都打上了标签，不仅描述该音频的乐器种类，还描述了该音频的演奏技巧。

然后，我们进行了有关中国传统乐器音频的乐器种类识别的研究。对于乐器种类的识别，我们提取了三种类型的特征，分别是频谱质心与频谱延展度，频谱对比度，梅尔频谱。通过实验发现梅尔频谱的效果最佳。同时我们基于 VGGNet 搭建了不同层数的卷积神经网络，最终发现 8 层的卷积神经网络识别效果最好。然后我们对数据集的划分进行了测试，发现我们的模型过拟合状况较小。最终我们使用梅尔频谱作为输入，训练 8 层的卷积神经网络进行识别，最终在测试集上取得了 99.3%的准确率，同时每一大类乐器之间几乎没有分类错误。

最后，我们进行了有关中国传统乐器音频的演奏技巧识别的研究。我们将演奏技巧的识别分为了两大部分，分别是单一乐器的演奏技巧识别和同一类乐器的演奏技巧识别。对于单一乐器的演奏技巧识别，由于其数据量较少，所以我们使用预训练的 ResNet 卷积神经网络模型提取梅尔频谱的特征，然后使用 SVM 进行分类识别。最终对于每一种乐器均取得了超过 99%的准确率。对于同一类乐器的演奏技巧识别，我们首先使用单一乐器的识别方法进行研究，四大类乐器的识别准确率如下：吹奏类 86.8%，弹拨类 66.5%，拉弦类 75.2%，敲击类 80.6%。然后我们自己搭建了卷积神经网络进行训练识别，四大类乐器的识别准确率如下：吹奏类乐器 95.7%，弹拨类乐器 82.2%，拉弦类乐器 88.3%，敲击类乐器 97.5%。

我们开放了中国传统乐器音频数据库和整个实验的 Python 源代码以供进一步研究。

5.2　展望

本文主要研究基于机器学习的中国传统乐器音频识别，为此专门构建了中国传统乐器音频数据库，同时基于该数据库我们完成了中国传统乐器音频的乐器种类识别和演奏技巧识别，目前还存在一些不足需要加以改进：

⑴针对中国传统乐器音频的乐器种类识别问题，我们主要研究了梅尔频谱特征，后续可以尝试提取其他类型的特征进行实验。同时论文中由于设备的限制我们只是基于

VGGNet 搭建了 8 层的卷积神经网络进行识别，后续可以尝试搭建多层的 ResNet 卷积神经网络来进行识别。

⑵针对中国传统乐器音频的演奏技巧识别问题，我们划分为了单一乐器演奏技巧识别和同一类乐器的演奏技巧识别。在实验中，我们均是采用有监督的学习。后续我们可以尝试对演奏技巧进行无监督的学习，包括传统的无监督学习和基于 GAN 神经网络的 ClusterGAN 无监督学习。

⑶本论文均是基于自己搭建的中国传统乐器音频数据库来进行实验的，为了检测实验中模型的鲁棒性，我们可以搜集网上的一些乐器音频，重新搭建一个测试音频数据集来检验模型的效果。

⑷本文主要是为研究中国传统乐器的各种课题做铺垫工作，属于基础性研究，为此本文开放了实验数据集和源代码。后续可以尝试多种乐器检测，乐器分离，乐器自动作曲等研究。

参考文献

[1] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5), 293-302.

[2] Deng, J. D., Simmermacher, C., & Cranefield, S. (2008). A study on feature analysis for musical instrument classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38(2), 429-438.

[3] Liu, J., & Xie, L. (2010, May). Svm-based automatic classification of musical instruments. In 2010 International Conference on Intelligent Computation Technology and Automation (Vol. 3, pp. 669-673). IEEE.

[4] Nagawade, M. S., & Ratnaparkhe, V. R. (2017, May). Musical instrument identification using MFCC. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 2198-2202). IEEE.

[5] Xu, C., Maddage, N. C., Shao, X., Cao, F., & Tian, Q. (2003, April). Musical genre classification using support vector machines. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). (Vol. 5, pp. V-429). IEEE.

[6] Li, T., & Ogihara, M. (2006). Toward intelligent music information retrieval. IEEE Transactions on Multimedia, 8(3), 564-574.

[7] Masood, S., Gupta, S., & Khan, S. (2015, December). Novel approach for musical instrument identification using neural network. In 2015 Annual IEEE India Conference (INDICON) (pp. 1-5). IEEE.

[8] 王飞. 基于音色分析与深度学习的乐器识别方法研究[D].江南大学,2018.

[9] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 13(2), 206-219.

[10] Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017, August). Timbre analysis of music audio signals with convolutional neural networks. In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 2744-2748). IEEE.

[11] Kong, Q., Iqbal, T., Xu, Y., Wang, W., & Plumbley, M. D. (2018). DCASE 2018 challenge surrey cross-task convolutional neural network baseline. arXiv preprint arXiv:1808.00773.

[12] 沈骏, 胡荷芬. 中国民族乐器的特征值提取和分类[J]. 计算机与数字工程, 2012, 40(09):119-121.

[13] 王芳. 基于深度学习的音乐流派及中国传统乐器识别分类研究[D]. 南京理工大学, 2017.

[14] Liang, X., Li, Z., Liu, J., Li, W., Zhu, J., & Han, B. (2019). Constructing a Multimedia Chinese Musical Instrument Database. In Proceedings of the 6th Conference on Sound and Music Technology (CSMT) (pp. 53-60). Springer, Singapore.

[15] Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., & Cai, L. H. (2002, August). Music type classification by spectral contrast feature. In Proceedings. IEEE International Conference on Multimedia and Expo (Vol. 1, pp. 113-116). IEEE.

[16] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.

[17] Smith, Julius O. Digital Audio Resampling Home Page Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, 2015-02-23.

[18] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[19] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

致　　谢

四年的大学生涯即将结束，明知道时光一去不复返，却依然缅怀过去。毕竟在这个过程中，我的心态，为人处事，思考方式等都发生了巨大的变化。在毕业设计论文即将完成时，我想向那些在一路中帮助过我的父母，老师，同学们和朋友们表达我最纯真的谢意！

首先，感谢我的导师李荣锋老师。从去年的题目申请到正式开题，从毕设中期答辩到即将进行的终期答辩，李荣锋老师都给予了我很大的帮助。每周一次的小组讨论让我及时找到自己的不足及改进措施，每次细致地沟通交流让我及时明白目前存在的问题及解决方案。感谢老师在课题研究方面给予的指导性意见和在论文撰写方面给予的改进建议。

其次，感谢整个毕业设计小组的同学们。每周一次的小组讨论不仅是分享自己的进展，同时也是学习他人的经验。在每一次的组会中，我不仅可以总结出自己项目的下一步计划，同时可以了解到大家的研究思路和研究方法，这对于我的灵感的产生起到了非常大的作用，感谢大家。

然后，感谢中国音乐学院提供音频数据库，我的所有研究都是在该数据库上进行的，所以感谢所有在数据库音频录制过程中贡献力量的朋友们。感谢我的父母，无论遇到什么问题，他们都会一直支持我。在考研失利时，不停奔波于找工作，找调剂，面试，笔试，做实验和撰写论文中，最后还是坚持了下来。

最后，感谢评审老师的审阅，谢谢！

附　　录

附录 1　中国传统乐器音频数据库

https://pan.baidu.com/s/1x3gAmabbIyJA9S_AeXYMvA 提取码：m117

附录 2　基于机器学习的中国传统乐器音频识别 Python 源代码

https://github.com/jinzhaochaliang/Chinese-Musical-Instruments-Classification.git

附录 3　图表索引

附录 4 公式索引