保密类别_____ 编　号_____

毕　业　论　文

中国民族乐器识别特征分析研究

学部(院)　音乐与录音艺术学院

专　　业　录音艺术（录音工程方向）

班　　级　2016 级录音工程

姓　　名　张辉程

指导教师　王鑫

中 国 传 媒 大 学

2020 年 5 月 9 日

中国民族乐器识别特征分析研究

张辉程

摘　要

乐器识别一直是音频信息检索（Music Information Retrieval）的重要任务。对于构建一个基于机器学习的乐器识别系统来说，特征提取和特征选择是关键的一步，尽管当下已经涌现了数以百计的前沿机器学习算法，比如说深度学习和迁移学习算法，在此类算法中不需要人工提取特征。有不少学者开展了基于西洋乐器的音频特征选择优化研究，但是关于中国民族乐器的音频特征选择研究不算太多。本文进行了中国民族乐器单音分类问题的特征选择研究，使用包裹式特征选择方法递归特征消除（Recursive Feature Elimination）来对预处理后的原始特征集进行特征选择和特征分析。实验结果证明，235 维原始特征集内部具有冗余性，最佳维度是 31 维；经过特征选择后，基于支持向量机（Support Vector Machine，SVM）分类模型的平均准确率（Accuracy）从 71%提升到 98%，得到了 27%的提高。

关键词：乐器识别，声学特征集，特征选择，ESSENTIA，PYTHON

A Study on Feature Analysis for Chinese Musical Instrument Indentification

Huicheng Zhang

ABSTRACT

Instrument identification is always an important task in music information retrieval. Feature extraction and selection is an essential part of building a machine learning project, even though hundreds of cutting-edge machine learning algorithms coming in these days like deep learning and transfer learning, which combine feature extraction in the learning process. Many fellows have conducted the study on feature optimization for western musical instrument identification, whereas those studies for Chinese musical instrument identification are rare. Thus, this essay presents feature extraction and selection for a large feature set extracted from samples in Database for Chinese Musical Instrument, using Recursive Feature Elimination method to generate the optimal feature set. After using this optimal feature set, the average accuracy achieved from Support Vector Machine classifier has increased from 71% to 98%, compared to the original feature set. Finally, feature analysis is carried out. In conclusion, it is necessary to implement feature selection within a machine learning project.

Keywords: Feature Selection, Chinese Musical Instrument Identification, Support Vector Machine

目 录

摘 要 I
ABSTRACT II
目 录 1
一、绪 论 2
（一） 本课题研究的背景和意义 2
（二） 国内外研究现状 2
（三） 本文主要工作与内容安排 5
二、乐器分类及实验数据库的建立 6
（一） 中国民族乐器分类 6
（二） 实验数据库概述 7
三、中国民族乐器识别特征提取和特征选择 9
（一） 音频特征概述 9
（二） 音频特征的分类及定义 9
1、时域能量包络特征 10
2、时域特征 10
3、能量特征 10
4、频谱包络统计特征 11
5、谐波特征 11
6、感知特征 12
7、倒谱域特征 13
（三） 特征集构建以及特征提取工具系统参数概述 13
四、中国民族乐器识别实验 16
（一） 数据预处理 16
1、数据标准化 16
2、数据分集 16
3、单变量分析 16
（二） 特征选择以及分析方法概述 17
（三） 乐器识别系统评价指标 17
（四） 特征选择实验以及特征分析 18
（五） 最佳特征集乐器识别实验结果分析 21
五、结论 26
参考文献 27
附录 29
后记 35

一、绪 论

（一）课题研究的背景和意义

声音有三要素：音调、响度和音色。根据美国国家标准机构的定义，音色是人们区别具有相同响度和音调的两个声音的主观感受。音色主要由声音的频谱结构决定，但同时，音色还与声音的时域包络、强度以及持续时间有关。心理声学实验表明，人对于乐器的音色感知和实际测量出来的频谱数据可能会有明显的差异。对于人类专家来说，通过音色分辨乐器是一件简单的事情，但其中涉及的机理却显得语焉不详。究其原因，是因为音色感知是一个较为复杂的心理现象。除此之外，建立在联觉基础上的音色评价术语具有一定的模糊性。上述事实对于在计算机上实现乐器识别增加了难度。

随着上世纪九十年代数字音频技术和互联网技术的迅猛发展，人们聆听音乐的方式发生改变，从听黑胶、听磁带，变成了聆听储存在电脑或者音频播放器上的数字化录音。这些庞大数量的音频文件促使了音频信息检索技术（Music Information Retrieval）的产生。

经过学者们多年的研究，使用数据科学、机器学习的方法来实现乐器识别成为可能。常见的工作流程是：对音频文件提取特征，使用特征来表征音频样本，然后使用机器学习算法来进行样本的分类识别。

乐器识别是音频信息检索领域的基础问题，在将近三十年的时间里，很多学者在西方乐器识别问题上做出了大量研究。但是中国民族乐器识别中的特征选择研究就显得相对有些匮乏。不少中国民族乐器识别研究是直接选用了西方乐器识别研究中常用的音频特征，尽管也获得了不错的识别准确率，但是，由于中西乐器间不同的声学构造和演奏风格，中国民族乐器的特征集应该会有不同的最优特征组合。因此进行中国民族乐器识别的特征选择和特征分析研究，可以帮助我们更进一步了解中国民族乐器的特性，为涉及中国民族乐器的其他 MIR 研究提供一些参考。

（二）国内外研究现状

特征选择方面的研究多见于西方学术圈：

Deng 等人[1]开展了西洋乐器识别中对于特征选择的研究。特征集包括基于听觉感知的特征、倒谱特征以及 MPEG-7 音频特征这三类特征共 20 个。首先通过信息增益

（Information Gain，IG）、信息增益比例（Gain Ratio，GR）、对称不确定性（Symmetrical Uncertainty，SU）指标，来检验特征的相关性并进行排序筛选，生成一个新的特征集。通过降维方法主成分分析（Principal Component Analysis，PCA）和等度量映射（Isomap）来探究特征选择的最佳特征数量。随后，特征集在三种乐器识别任务（乐器族归类、单件乐器识别、独奏乐段识别）中通过 k-NN、朴素贝叶斯、多层感知机（Multilayer Perceptron，MLP）、径向基函数（Radial Basis Functions，RBF）和 SVM 分类模型进行进一步测试。实验结果表明 MPEG-7 中的对数起振时间（Log Attack Time，LAT））和谐波谱偏差（HD，Harmonic Deviation）以及梅尔频率倒谱系数（MFCC）的表现较为突出，另外也证明了特征集中具有冗余性，特征选择对于优化分类任务中的特征集是有必要的。

Mingchun Liu 等人[2]为西洋乐器数据集（共 351 个单音文件）提取了 58 个特征，然后使用序列前向选择方法来选择最佳特征集，以此获得较高的乐器识别准确率。实验同时测试了三种不用的分类模型（NN，modified k-NN，GMM），结果表明使用 19 个特征可以获得最高为 93%的识别准确率。

Uruthiran 和 Ranathunga[3]使用序列前向选择方法以及贪婪搜索算法（The Greedy Search）来寻找最佳特征集。数据库包括 13534 个来自爱乐乐团的乐器单音样本，涵盖弦乐、铜管、木管三种乐器组共 20 种乐器。通过 Matlab 提取特征构建特征集，其中包括频域、时域和倒谱域特征共 44 维。实验开展时，特征的数量从 2 维逐一增加到 44 维，当特征集的维度到 19 维的时候，分类模型可达到最高的准确率 91.43%。随后基于最优特征集开展乐器识别实验，同时对比了 SVM，kNN 和决策树这三个分类算法的效果。实验表明，对于 95%的乐器识别案例，SVM 都能得到最高的识别准确率。这说明 SVM 是针对该最优特征集最为合适的识别分类算法。

后方帅[4]研究 13 种西洋管弦乐器单音识别，特征选择阶段使用基于 Fisher 准则的特征选择和基于信息增益（Information Gain, IG）的方法，对由 Timbre ToolBox 提取的时域频域特征以及倒谱域特征所组成的 148 维特征集进行特征选择。实验发现，按照 Fisher 准则的特征选择处理后，最优混合成份数量以及特征维数分别是 256 和 50，所得到的特征集整体准确率仅为 78.6%；而基于 IG 准则的特征选择处理后得到的最优参数分别是 8 和 30，所得到的特征集整体准确率是 90.1%。对比两种方法选择出的音色特征的整体准确率可知，IG 准则更善于锁定更高性能的音色特征。

除了特征选择研究，更多学者着力于特征提取和分类算法的研究上：

谢凌云[5]利用自建数据库（包含中西方乐器，共 2177 个时长为 30 秒的乐器音乐独奏片段），提取 STFT（短时傅立叶变换特征，包括频谱质心、谱下降值、谱通量）、MFCC、谱峰值因子 SCF、谱平整度 SFM 四种特征子集作为 WEKA 分类算法模型的输入，实验结果发现对于中国民族乐器，SFM 比 MFCC 更适合于中国乐器识别，SFM 搭配 KNN 分类算法是识别准确率最高的组合。

刘璇、许洁萍、范丽[6]开展了的乐器识别研究中考虑到音频特征的动态特性，论文基于调制谱的思想实现了三个长时特征的提取，分别是多变量自动回归模型（Multivariate Auto Regressive, MAR），动态 Mel 频率倒谱系数（Dynamic MAR, DMFCC），音频子带时域包络(Auditory Filterbank Temporal Envelopes, AFTE)多变量自动回归模型（Multivariate Auto)。论文对 11 种短时特征（共 72 维）和三种长时特征进行组合实验，分析结果得出 8 类乐器（琵琶、二胡、萨克斯、钢琴、笛子、小提琴、吉他、古筝）与各特征集之间的分类识别偏好，并利用多特征组合分类的思想结合乐器识别偏好结果，提出了针对乐器识别任务的特征组合算法。实验结果验证了此算法的有效性，8 类乐器识别准确率可以达到 86.34%。

杨靖[7]为乐音提取离散谐波变换系数（DHTC）以及它们的一阶、二阶差分，共同构建音色表达谱，同时提取时频倒谱域特征（38 维时域特征、44 维频域特征、12 维线性预测倒谱系数 LPCC、23 维 MFCC 及其一阶差分）。实验发现，选择音色表达谱和 LPCC、MFCC 组成的 53 维特征集，使用支持向量机作为分类器，在乐器单音数据库上对 25 类乐器的识别率达到 90.2%。

龙丽婧[8]对基于时间变化的音色进行改进，通过检测短时状态边界来实现对音色识别精确度的提高。首先为不同长度的音频素材设置不同的瞬时边界，然后提取特征（包括谱分布方差、谱质心等），最后由决策树-J48 和 k-NN 分类算法建立乐器分类模型。结果发现，所加入的改进特征对个别乐器识别有恶化作用，但对整体平均识别准确率有改善作用。

王琪[9]以 MFCC 为基础进行西方乐器的音色识别研究。提出基于非音高成分的梅尔频率倒谱系数（NMFCC）以及基于低阶本征模态的梅尔频率倒谱系数（LOMS-MFCC）。先使用经验模态分界（Empirical Mode Decomposition, EMD）对信号做平稳化处理，通过提取出隐藏在信号中的振荡模态，最终将不同时间尺度的信号分量区分开来。分解后得到的每一个振动模态称为本征模态函数（Intrinsic Mode Function，IMF）。实验证明，NMFCC 和 LOMS-MFCC 对于弦乐同族乐器的识别效果非常好。

Shubham Bahre 等人[10]使用了由起振斜率、常数 Q 变换系数和倒谱特征组成的 261 维特征集，基于 SVM 算法对五种西方乐器进行单音识别，获得了平均 86.75%的准确率。

还有部分学者研究音频特征之间的统计特性差异：Gulhane Sushen R.等人[11] 通过统计数据分析方法进行了关于乐器识别中倒谱特征和频谱特征的定量研究。根据特征对于不同西方乐器和印度乐器的的差异程度，研究推测 MFCC 系数 1 和频域特征中的谱滚降和谱质心足以完成乐器识别任务。

刘若伦，张家琦[12]从激励源-共振体的乐音模型出发，针对西洋乐器木管乐器族内各个乐器音色相似的问题，他们认为要消除共振体的影响，因此使用点阻滤波器对各谐波进行调整，使其幅度相同。经过点阻滤波的乐音信号称为机制声信号。他们提出新特征机制声 MFCC,通过观察 MFCC 和机制声 MFCC 每个因素的数值在不同乐器上的

统计差异，他们发现各个乐器的机制声 MFCC 更容易被区分。

通过对以往乐器识别研究的整理，可以看到国内外已存在大量乐器识别的研究，多数学者尝试提出新的特征集或者使用新的分类算法，然后通过乐器识别实验来验证该特征或者该分类器性能的优劣，最后可以得出特征集和分类器的最佳搭配。而对于特征选择研究来说，目前针对中国民族乐器识别的特征选择研究还比较少，因此，本文打算针对中国民族乐器识别特征选择开展研究，寻找适用于中国民族乐器识别的特征集。

（三）本文主要工作与内容安排

本论文的主要工作是：参考其他学者的特征分析研究，整理出一个比较全面的用于乐器单音识别的特征分类框架，并利用 Timbre Toolbox、Essentia 以及 Librosa，对 26 件中国民族乐器共 5654 个单音的数据库构建出 235 维初始特征集。随后使用包裹式特征选择方法——递归特征消除（RFE），来得出最佳特征集的维度数、同时完成最佳特征集的构建。最后将最佳特征集用在基于支持向量机的分类模型训练上，并对比特征选择前的分类情况，来考察最佳特征集的性能。

本论文的内容布局如下：

第一章介绍国内外对于乐器识别以及特征选择的研究现状；

第二章介绍乐器分类和实验数据库的基本情况；

第三章介绍音频特征的分类以及原始特征集的构建；

第四章介绍特征选择实验和中国民族乐器识别实验；

第五章为结论部分。

二、乐器分类及实验数据库的建立

（一）中国民族乐器分类

1914 年，萨克斯（C. Sachs）与霍恩博斯特尔（E. Horhbostel）在其著作《乐器分类法》[13]中将乐器进行多级分类。第一级分为体鸣、膜鸣、弦鸣、气鸣和电鸣乐器五大类，随后又按演奏方式、乐器体制等标准进行逐级细分。这种分类方法具有科学性、逻辑性。而在管弦乐队分类法中，乐器则被划分为弦乐器、管乐器、打击乐器三大类，这是针对管弦乐的音乐实践，分类的结果和乐队的声部保持一致，相对于萨克斯的分类方法更具有实用性。

中国历史上关于乐器分类方法最早基于制作材料的，早在周朝已有记载，称为“八音分类法”，将乐器分为“金、石、土、革、丝、木、匏、竹”。这种分类方法与非音乐文化具有更多的关联（比如汉字、风水），在音乐实践的过程中显得不太实用，对于非上述材料制作的乐器无法进行分类。二十世纪六十年代初，中国音乐学家结合我国数千年的器乐发展史，以乐器演奏方式为出发点，按照乐器的形状和性能将中国传统乐器分为拉弦乐器、弹拨乐器、吹奏乐器、打击乐器四大类。这种分类方式以人的演奏动作作为分类标准，体现了人的主体性；不同于西方管弦乐队分类法，上述分类方法将弹拨乐器划分为一级类别，这突出了鲜明的民族性，同时在民族管弦乐的实践中也十分实用。本文将沿用这种乐器分类方法。

（二）实验数据库概述

本实验基于 DCMI（Database of Chinese Musical Instrument）数据库建立的。DCMI 数据库[14]是由中国音乐学院音乐科技系于 2018 年组织建立的一个关于中国民族乐器的多媒体数据库，网站（http://47.90.12.82）包含关于乐器的文字描述、图片、乐器声学测量报告、乐器的录音、关于乐器制作流程的视频。录音工作在全消声室内进行，预计录入 200 个具有代表性的乐器，主要是由民族管弦乐器以及中国戏曲伴奏与乐器组成。DCMI 数据库内的乐器录音包含空弦音（弦乐器）、最低音（管乐器）以及不同的力度（强奏、中强、弱奏）、不同的音阶（半音阶、五声音阶、自然大调音阶）、不同的演奏技巧（正常演奏技巧、特殊演奏技巧）以及音乐片段。

本论文研究中国民族乐器的单音识别，因此需要对 DCMI 数据库的录音文件进行单音提取。这部分工作是通过在 Jupyter Notebook 中编写脚本，利用 Python 工具库 Pydub 批量完成对音频文件内静音段的检测和分隔工作。计算机生成单音文件后，再

人工逐一检查样本，对于个别分隔不当的样本进行手动切割，所使用的音频编辑软件是 Adobe Audition CC。

对于特殊技巧音色的筛选标准是：仅保留单次激励的、具有主体音高的特殊技巧音色。也就是说，多次运弓、多次拨弦、多次吐气和多次敲击的音色不在考虑范围，同一时间出现多个音高的音色也被排除在外（拉弦乐器的“双音”，拨弦乐器的扫弦等等）。

按照以上标准总共产生的样本数量为 5654 个、涵盖乐器种类共 26 种，音频格式均为单声道 wav 格式、采样频率为 44.1kHz，量化深度是 16 比特。详细的实验数据库情况如下表所示。

表格 1 实验数据库情况

乐器分类		乐器名称（后缀为 DCMI 数据库中的编号信息）	常规音色	特殊技巧音色	样本数量	
弦乐器	拉弦乐器	二胡 L0266	649	221	870	1590
		中音板胡 L0240	192	64	256	
		高音板胡 L0239	133	64	197	
		椰胡 L0288	80	0	80	
		六角高胡 L0292	187	0	187	
	拨弦乐器	三弦 T0289	653	97	750	2626
		琵琶 T0262	463	32	495	
		中阮 T0260	720	0	720	
		古筝 T0255	42	38	80	
		柳琴 T0261	401	94	495	
		筚篥 T0254	76	10	86	
吹奏乐器		箫 C0282	54	25	79	1280
		A 调曲笛 C0280	71	28	99	
		G 调新笛 C0281	54	28	82	
		G 调梆笛 C0237	216	105	321	
		葫芦丝 C0309	52	151	203	
		传统笙 C0244	180	70	250	
		埙 C0283	25	16	41	
		唢呐 C0296	150	63	213	
打击乐器		大镲 D0290	6	8	14	150
		云锣 D0279	58	0	58	

	中国大鼓 D0248	6	17	23	
	小堂鼓 D0250	6	6	12	
	小镲 D0271	4	8	12	
	铙 D0269	6	9	15	
	铙钹 D0270	6	10	16	
乐器数量总和		5654			

三、中国民族乐器识别特征提取和特征选择

（一）音频特征概述

音频特征的诞生起源于一系列关于音色感知的研究。“音色”这一个属于感知领域的术语与其他声音感知要素密切相关。许多学者研究乐音音色模型的构建，他们把音色定义为一个多维度现象，并使用“音色空间”（Timbre Space）来表征音色的感知结构。

从音频信号中提取出与上述感知维度有所关联的“声学参数”是一项重要的工程，这种“声学参数”称为“音频特征”（Audio Descriptor）。Grey 和 Gordon[15]较早开始尝试对声音信号在感知维度和物理声学维度之间的关联程度进行定量研究，谱质心（Spectral Centroid）是他们的研究结果。随后其他西方学者的研究更为全面系统地进行阐述音色空间中所有感知维度与音频特征的相关性。

近年来，有许多学者着力研究对于中国民族乐器的音色模型问题。Wei Jiang 和 Jingyu Liu 等人[16]通过主观评价实验构建一套评价乐器音色的术语系统，然后使用多维标度分析（Multidimensional Scaling ,MDS）来对主观评价实验中得到的差异矩阵（Dissimilarity Matrix）进行降维，再使用线性回归、支持向量回归、神经网络以及随机森林算法，来对每个感知维度与 166 维音频特征集之间的相关性进行分析，最后建立了 37 件中国民族乐器的 3D 音色空间模型。

在机器学习领域中，音频特征是对原始信号的精简表征。这个特点具有天然的内在矛盾，即特征需要尽可能小、但同时也要尽可能完整地保留信号的特点。选取合适的音频特征有助于对机器学习算法的效率提升。

（二）音频特征的分类及定义

本论文主要参考了 Timbre Toolbox 的特征分类方法[17]，并结合本实验使用到的工具进行了个别特征的补充。特征提取算法的处理对象是五种音频信号表征（Representations of the audio signal），它们分别是原始音频信号$s(t_n)$及其经过四种变换后的信号，四种变换为：1、希尔伯特变换（Hilbert Transform）；2、短时傅立叶变换（STFT）；3、听觉模型（Auditory Model），听觉滤波器有如等效矩形带宽（Equivalent Rectangular Bands，ERB）、梅尔频带（Mel Bands）、巴克频带（Bark Bands）；4、正弦谐波模型（Sinusoidal Harmonic Partials）。基于上述四种音频信号表征，音频特征提取工具可产生了七种类别的音频特征，如下文所示：

1. 时域能量包络特征

此类特征是针对原始音频信号经过希尔伯特变换产生的能量包络信号$e(t_n)$进行计算的。比如：对数起振时间（Log Attack Time，LAT）、时域质心（Temporal Centroid，TC）、起振时间（Attack-time）、时域起振衰减（Temporal increase or decrease）、有效时长（effective duration）。Timbre Toolbox 中采用 Weakest-effort 方法来对信号的音头进行估算，这比以往采用固定阈值来估算音头开始时间节点t_{st}和音头结束时间节点t_{end}更具有鲁棒性。

对数起振时间可用如下公式表示：

$$LAT = \log_{10}(t_{end} - t_{st}) \tag{3-1}$$

时域质心是指能量包络的质心。它能够区分乐音中的瞬态成分和稳态成分。计算公式如下：

$$tc = \frac{\sum_{n=n_1}^{n=n_2} t_n \cdot e(t_n)}{\sum_n e(t_n)} \tag{3-2}$$

其中n_1和n_2分别是使得$e(t_n)$到达峰值 15%的首个和最后一个值，

2. 时域特征

时域特征是直接对原始音频信号进行计算而得，如自相关系数（Autocorrelation Coefficients）和过零率（Zero Crossing Rate）。自相关系数计算的是信号于自身不同时间点的互相关，它可以反映时域信号的频域结构，比如用来识别噪声掩蔽下的周期信号。Timbre Toolbox 中只保留前 12 个系数（$c \in \{1,\dots,12\}$），令L_n为窗口宽度（单位是采样点数），c为自相关的时间延迟。计算公式如下：

$$xcorr(c) = \frac{1}{xcorr(0)} \sum_{n=0}^{L_n - c - 1} s(n)s(n+c) \tag{3-3}$$

过零率计算的是单位时间内信号改变符号的次数，然后除以帧大小，简单来说是单位时间内信号在时域上通过零轴的次数。物理意义上过零率与信号频率一定程度上存在相关性，另外在实际情况中，不同种类信号的过零率大小关系通常为：空隙噪声 > 语音，音乐信号 > 语音信号。因此可以广泛用于语音活动（Speech Activity）的检测和音频分类。对于长度为N的音频帧，过零率的计算公式如下：

$$Z(i) = \frac{1}{2N} \sum_{n=1}^{N} |sgn[x_i(n)] - sgn[x_i(n-1)]| \tag{3-4}$$

3. 能量特征

通过不同音频信号表征所提取的能量特征。比如：基于短时傅立叶变换计算的频

域能量以及基于正弦谐波模型计算的谐波能量（Harmonic Energy）、噪声能量（Noise Energy）。在计算能量的过程中，窗函数的幅度经过了归一化，因此窗函数的种类和长度并不会影响计算结果。

4. 频谱包络统计特征

这类特征是针对频谱包络信号（可以基于短时傅立叶变换和正弦谐波模型）及其通过听觉滤波器修饰后的频谱包络信号进行统计计算而得。比如：谱质心（Spectral Centroid）、谱分布方差（Spectral Spread）、谱偏态（Spectral Skewness）、谱峰度（Spectral Kurtosis）、谱通量（Spectral Variation、Spectral Flux）。

以下介绍关于频谱统计特征的计算方法。令$\alpha_k(t_m)$代表t_m时刻音频帧中频点k所对应的短时傅里叶变换幅度。定义f_k为频点k对应的频率。再定义α_k的归一化形式$p_k(t_m) = [\alpha_k(t_m)]/\sum_{k=1}^{K}\alpha_k(t_m)$。

谱质心指的是频谱的重心。计算公式如下：

$$\mu_1(t_m) = \sum_{k=1}^{K} f_k \cdot p_k(t_m) \tag{3-5}$$

谱分布方差是指谱分布在其中心值（谱质心）的分布情况，描述了谱分布相对于质心的离散程度。计算公式如下：

$$\mu_2(t_m) = \left(\sum_{k=1}^{K}\left(f_k - \mu_1(t_m)\right)^2 \cdot p_k(t_m)\right)^{1/2} \tag{3-6}$$

谱偏态用来度量谱分布对于谱质心的对称性。$\mu_3 = 0$意味着对称分布，$\mu_3 < 0$说明更多能量分布于谱质心之下，$\mu_3 > 0$说明更多能量分布在谱质心之上。计算公式如下：

$$\mu_3(t_m) = \left(\sum_{k=1}^{K}\left(f_k - \mu_1(t_m)\right)^3 \cdot p_k(t_m)\right)/{\mu_2}^3 \tag{3-7}$$

谱峰度用来衡量频谱在谱质心附近分布的平坦程度。$\mu_4 = 3$ 是正态分布，$\mu_4 < 3$说明是更为平坦的分布，$\mu_4 > 3$说明是更为集中的分布。计算公式如下：

$$\mu_4(t_m) = \left(\sum_{k=1}^{K}\left(f_k - \mu_1(t_m)\right)^4 \cdot p_k(t_m)\right)/{\mu_2}^4 \tag{3-8}$$

谱通量用来衡量频谱随时间的变化。定义为 1 减去相邻α_k的归一化自相关。计算公式如下：

$$Variation(t_m, t_{m-1}) = 1 - \frac{\sum_{k=1}^{K}\alpha_k(t_{m-1})\alpha_k(t_m)}{\sqrt{\sum_{k=1}^{K}\alpha_k(t_{m-1})^2}\sqrt{\sum_{k=1}^{K}\alpha_k(t_m)^2}} \tag{3-9}$$

5. 谐波特征

此类特征是通过信号的正弦谐波模型计算而得。正弦谐波模型，或者叫正弦波-噪声模型（Sinusoidal plus Residual Model），由 Serra 和 Smith[18]在 1990 年所提出，他将乐音看成是随时间变化的正弦波和受调制白噪声的合成信号，主要目的是还原频谱中的周期性谐波信号和非周期性噪声信号，受调制的包络可以保持原乐音的时间特性。正弦谐波模型尝试在频谱角度去对耳膜接收到的声音信号进行模拟，只关注

人耳能够接收到的信号。相关特征有：谐波噪声比例（Harmonic/Noise Ratio）、奇偶谐波比例（Odd-to-Even Harmonic Ratio）、谐波三刺激值能量比例（Harmonic Tristimulus）、不和谐度（Inharmonicity）以及谐波偏差（Harmonic Spectral Deviation）。

谐波三刺激值由 Pollard 和 Jansson 提出，借鉴了视觉上三原色的概念。 它包含三个能量比例值，提供了频谱低次谐波结构的信息。计算公式如下：

$$T1(t_m) = \frac{\alpha_1(t_m)}{\sum_{h=1}^{H}\alpha_h(t_m)} \quad (3\text{-}10)$$

$$T2(t_m) = \frac{\alpha_2(t_m)+\alpha_3(t_m)+\alpha_4(t_m)}{\sum_{h=1}^{H}\alpha_h(t_m)} \quad (3\text{-}11)$$

$$T3(t_m) = \frac{\sum_{h=5}^{H}\alpha_h(t_m)}{\sum_{h=1}^{H}\alpha_h(t_m)} \quad (3\text{-}12)$$

其中H是所研究的谐波总数量（在 Timbre Toolbox 中默认参数为$H = 20$）。

不和谐度测量的是所测分音f_h（Partials）与基频整数倍hf_0间的偏离度，通过计算每一个分音偏离值的加权求和而得。计算公式如下：

$$inharmo(t_m) = \frac{2}{f_0(t_m)}\frac{\sum_{h=1}^{H}(f_h(t_m)-hf_0(t_m))\alpha_h^2(t_m)}{\sum_{h=1}^{H}\alpha_h^2(t_m)} \quad (3\text{-}13)$$

谐波谱偏差测量的是所有分音对于全局频谱包络$SE(f_h,t_m)$的平均偏离度，它反应了谐波包络的平滑程度。计算公式如下：

$$HDEV(t_m) = \frac{1}{H}\sum_{h=1}^{H}(\alpha_h(t_m) - SE(f_h,t_m)) \quad (3\text{-}14)$$

其中$SE(f_h,t_m)$代表了在频点f_h和时间t_m的频谱包络。在频点f_h的频域包络计算方法是计算相邻两个频点和自身的平均值。计算公式如下：

$$SE(f_h,t_m) = \frac{1}{3}\left(\alpha_{h-1}(t_m) + \alpha_h(t_m) + \alpha_{h+1}(t_m)\right), \quad 当\ 1 < h < H时 \quad (3\text{-}15)$$

奇偶谐波比例主要应用在区分不同谐波能量分布的乐器，比如双簧管奇次谐波占主要能量、频谱结构比较参差不齐，小号的频谱结构则更为平滑。计算公式如下：

$$OER(t_m) = \frac{\sum_{h=1}^{H/2}\alpha_{2h-1}^2(t_m)}{\sum_{h=1}^{H/2}\alpha_{2h}^2(t_m)} \quad (3\text{-}16)$$

6. 感知特征

此类特征是通过听觉模型的处理计算而得。比如平均响度（Average loudness）、相关标准响度（Related Relevant loudness）、不协和度（Dissonance）和动态复杂度（Dynamic Complexity）。

平均响度是基于心理学家 Stevens 幂定律进行计算的，算法认为心理量“响度”等于刺激量“音频信号能量”的 0.67 次幂。相关标准响度是根据欧广联响度 R128 标准进行计算的。音频信号先通过 K 计权曲线进行滤波处理。瞬时响度（Momentary Loudness）是通过对窗长为 400 毫秒的矩形窗内的信号能量进行求和，然后再求统计

量。短时响度（Short Term Loudness）是通过对窗长为 3 秒的矩形窗内的信号能量进行求和，然后再求统计量。整合响度（Integrated Loudness）的测量需要经过两个特殊门限进行计算。

不协和度是基于 Plomp 和 Levelt 的“临界频带”理论[19]进行计算的。首先计算出音频信号谱峰（Spectral peaks）的频率和幅度，然后按照感知实验提出的不协和曲线得出每一对谱峰的不协和度，最后将每对谱峰归一化后的不协和度值进行求和。

动态复杂度计算的是全局响度分贝值的平均绝对偏差（Average Absolute Deviation）。动态复杂度与录音的动态范围有关系，为了保证结果的准确性，算法忽略了音频文件首尾的静音段。

7. 倒谱域特征

此类特征基于音频信号的频谱包络计算而得。比如：梅尔频率倒谱系数（Mel-frequency cepstral coefficients，MFCC）及其一阶差分（delta Mel-frequency cepstral coefficients， DMFCC）、二阶差分（delta delta Mel-frequency cepstral coefficients，DDMFCC）。梅尔频率倒谱系数是通过信号在梅尔频率刻度（Mel Scale）频域进行离散余弦变换计算而得。基于听音实验得出的梅尔频率刻度中，相邻的频率在听觉上音程距离相等，它是一个非线性尺度，因其与人耳听觉规律接近而被广泛使用。梅尔频率F_{mel}与信号频率f的对应关系是：

$$F_{mel} = 2595\log_{10}(1+\frac{f}{700}) \qquad (3\text{-}17)$$

其中梅尔频率F_{mel}的单位是美尔，信号频率f的单位是赫兹，基准参照点是 1000Hz 的纯音等于音高 1000Mel。计算步骤如下：

1、原始音频信号s；

2、对s分帧、加窗，得到音频帧；

3、对于每个音频帧通过快速傅立叶变换 FFT 得到线性频谱功率幅度α_s；

4、α_s 通过 Mel 滤波器组，并将每个滤波器中的能量进行求和；

5、对通过该滤波器组后频谱能量进行取对数；

6、经过离散余弦变换 DCT 求得正交化后的 MFCC 系数；

7、在 24 个系数里取能量最集中的前 12 个或者 13 个作为最终结果。

（三）特征集构建以及特征提取工具系统参数概述

本实验为了尽可能全面地构建特征集，选用了三种特征提取工具。特征提取工作主要在 Matlab 的工具包 Timbre Toolbox 中完成，另外使用 Librosa 与 Essentia 分别进行倒谱域特征以及感知特征的补充提取。机器学习部分则利用 Python 工具库 Scikit-learn，来完成特征选择、构建并训练乐器分类模型的工作。

以下是本实验所构建的特征集。如下表所示，完整表格请看附录。

表格 2 实验特征集

特征类别	维数
能量特征	17
时域能量包络特征	8
时域特征	26
频域频谱包络统计特征	42
谐波特征	17
谐波频谱包络统计特征	16
感知特征	10
听觉模型频谱包络统计特征	60
倒谱域特征	39
总特征数	235

特征提取工具参数如下文所示：

时域能量包络是在希尔伯特变换后的信号中提取出来的。幅度信号随后被一个截止频率为 5 赫兹的三阶巴特沃斯滤波器进行低通处理，形成时域能量包络。

短时傅里叶变换使用汉明窗（Hamming），窗口宽度是 23.2 毫秒，帧移（Hop size）是 5.8 毫秒。短时傅里叶变换有两种尺度，一种是幅值（Magnitude），另一种是功率（Power）。

在 Timbre Toolbox 中，听觉模型是指等效矩形带宽，它是通过一系列带通滤波器来模拟人类听觉的掩蔽效应，这与“临界频带”的概念是类似的。Timbre Toolbox 中提供两个方案，一个是使用 Gamatone 滤波器，随后进行时域平缓（temporal smoothing），另一个方案是基于 STFT。

正弦谐波模型使用布莱克曼窗（Blackman），窗函数大小为 100 毫秒，帧移是 25 毫秒。窗函数的大小比 STFT 计算的要大，这是为了获得更好的频谱分辨率。

对于帧级别特征（Time-varying audio descriptor）的统计方法，Timbre Toolbox 只使用中位数（Median）来度量数据中心以及使用四分位距[1]（Interquartile Range，IQR）来度量数据的分散情况。因为在针对音频信号进行特征提取时，所提取的特征有可能是针对静音帧的。这些静音帧的特征值是异常值（Outlier），并对于统计计算平均值、标准差、最小值产生严重的坏影响。而如果使用一个基于响度的门限对乐音信号进行静音段去除预处理，门限的定量设置会成为一个问题，不利于特征提取算法在不同数据集、在不同种类的音频信号上的适用性。所以 Timbre Toolbox 选择计算中位数和四分位距这些鲁棒性更高的统计手段。

[1] 四分位距是指经过大小排列的数据中，第三个四分位数（75%）和第一个四分位数（25%）的差值，与标准差一样，表示统计数据中各变量的分散情形。

为补充听觉模型特征以及感知特征，另外使用 Essentia 进行特征提取。Essentia 可以对信号功率频谱经过巴克滤波器组和梅尔滤波器组处理后的信号进行频谱包络统计特征的提取，还能对音频信号进行 EBU128 标准响度的计算。对于底层特征的提取，默认参数帧长（FrameSize）等于 46.4 毫秒、帧移等于 23.2 毫秒、窗函数是布莱克曼-哈里斯窗（Blackman-Harris）。在计算响度时，帧长默认等于两秒，帧移等于一秒。本实验系统对于 Essentia 所提取的帧级别特征采用的统计手段，是中位数和标准差。

为了提取 MFCC 及其一阶差分、二阶差分，本系统使用 Python 工具库 Librosa 进行特征提取。值得注意的是，Librosa 默认的采样频率是 22050Hz，MFCC 的数量是 20。为了和本系统其他特征提取工具算法的参数保持一致，需要把采样频率设置成 44100Hz、把 MFCC 提取数量设置成 13。与上述特征提取工具对于帧级别特征的统计方式不同，本实验中所有倒谱域特征是采用平均值进行统计。

四、中国民族乐器单音识别实验

本系统主要的编程语言是 Python，在 Jupyter Notebook 中进行特征提取、特征选择以及乐器识别实验的编程设计。所有代码根据不同的功能编写在多个 Jupyter Notebook（.pynb）文件中。交互性的开发环境可以直观地看到程序的结果，在调试的过程中显得更加方便。

本实验为 5654 个单音样本都提取了 235 个特征，在编程环境中形成了一个 5654*235 的二维数组，一个特征指的是二维数组中的其中一列。不一定所有列都会与输出变量（乐器标注）有强相关性，如果把不相关的特征加入到模型里，机器学习模型的输出会很差。特征选择的目的是生成一个精简的最佳特征集，以便后续开展乐器识别研究。

（一）数据预处理

1. 数据标准化

使用 Scikit-learn 模块中的 StandardScaler 来对数据进行标准化处理，使其平均值为零、方差为一，服从正态分布。对于机器学习算法来说，标准化数据是运算的假设对象。如果数据不大致符合正态分布，学习器的表现会大打折扣，所以有必要对特征集进行标准化处理。

在 Scikit-learn 中实际的做法是将每一维度的特征减去其平均值μ，以便将数据分布在中心，随后通过除以特征值的标准差σ来实现归一化。计算方式如下：

$$x^* = \frac{x-\mu}{\sigma} \qquad (4\text{-}1)$$

2. 数据分集

将 25%的样本随机抽取出来作为测试集，在特征选择过程中，测试集样本不会被处理，所有操作是基于训练集。控制随机抽取的参数 Random state 设置为一个确定的数 1，这样可以确定每次运行数据分集程序得到的样本一样，便于对特征选择算法参数的调试。

3. 单变量分析

单变量分析将待测变量与目标变量进行比较测试，比较的时候忽略除待测变量外的其他变量。测试分数越高，该变量与目标变量的关系越紧密。常见的测试函数有 chi-square（chi^2）和 F-regression，本研究采用 F-regression 方法。选取测试分数较高的 150 个特征进行下一步的研究，剔除了约 36%的特征（共 85 个）。因为实验过程中曾尝试依次改变单变量分析参数，使得输出的最优特征数量为 100 和 200 个，结果发现 RFE 的计算结果最优特征数量不尽相同，分别是 65 维和 49 维，对应的识别准确率分别为 94.06%和 96.18%。这说明尽管有些变量在单变量分析中被去除，但是

它们对识别准确率有些许提升作用。出于折中考虑，本实验设定单变量分析的初步筛选的特征数量是 150 个。

（二）特征选择以及分析方法概述

特征选择的意义，不仅可以消除不相关的特征来减少运算和储存损耗，还可以选出一个能够提高准确率的特征子集。特征选择方法有三种，过滤式（Filter）、包裹式（Wrapper）和包裹式（Embedding）。

过滤式方法先对先对数据集进行特征选择，然后训练学习器，特征选择过程与后续学习器无关。这相当于先用特征选择过程对初始特征进行“过滤”，然后过滤后的特征被用来训练模型。Relief（Relevant Features）是一种著名的过滤式特征选择方法，也可以通过计算特征之间的皮尔逊相关性来完成特征选择。

包裹式特征选择直接把最终将要使用的学习器的性能作为特征集的评价准则，也就是说，包裹式特征选择的目的就是为了给定学习器选择出最有利于其性能的特征集。有很多种特征选择方法，如：后向消除（Backward Elimination）、前向选择（Forward Selection）、双向消除（Bidirectional Elimination）和递归特征消除（Recursive Feature Elimination）。

嵌入式特征选择是将特征选择过程与学习器训练过程融合为一体，两者在同一个优化过程中完成，即在学习器训练过程中自动地进行特征选择。典型的嵌入式特征选择方法是使用 Lasso 回归模型进行特征选择。

（三）乐器识别系统评价指标

乐器识别系统的评价指标有精确率（Precision）、召回率（Recall）和 F1-score。

在二分类问题中，可以将样例根据其真实类别与学习器预测类别的组合划分为真正例（True Positive）、假正例（False Positive）、真反例（True Negative）、假反例（False Negative）四种情形。令TP、FP、TN、FN分别表示其对应的样例数，则有$TP+FN+FP+TN=$样本总数。由此可得出准确率定义：

$$Accuracy=\frac{TP+TN}{TP+FN+FP+TN} \tag{4-2}$$

混淆矩阵（Confusion Matrix）如下表所示。

表格 3 二分类问题混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

精确率也叫正预测值，指系统反馈的正例样本中准确正例所占的比例，体现了系统反馈结果的有用程度。即：

$$Precision = \frac{TP}{TP+FP} \quad (4\text{-}3)$$

召回率（也叫敏感度 Sensitivity），是指所有相关例子中被系统正确选中的比例，也就是预测的准确正例在全部真实正例中的比例,即：

$$Recall = \frac{TP}{TP+FN} \quad (4\text{-}4)$$

对于复杂的分类问题，优秀的精确率和召回率往往不可兼得。因此在二元分类问题的统计分析中，常常用 F1-score 来测量分类器的效率。F1-score 是召回率和精确率两个指标的加权调和，可表示为：

$$F1 = \frac{(1+\alpha^2)\cdot P\cdot R}{\alpha^2 P + R} \quad (4\text{-}5)$$

其中，α度量了精确率和召回率的相对重要性（$\alpha > 0$）。$\alpha > 1$时召回率更重要，$\alpha < 1$时精确率更重要，$\alpha = 1$时公式退化为平衡点（Break Event Point， BEP）的定义，此时两者同样重要。在本实验中，α设置为 1。

（四）特征选择实验以及特征分析

本实验采用包裹式特征选择方法递归特征消除（Recursive Feature Elimation，RFE）来进行特征选择，由于后续乐器识别系统的分类模型是基于支持向量机构建起来的，因此这里采用 SVM 作为学习器。内核函数是线性函数（Linear）。将识别准确率（Accuracy）作为特征选择指标，通过交叉验证（Cross validation）的方法来找到最佳特征集。实验表明，最佳特征集大小为 31 维，此时特征集识别准确率为 95.05%。如下图所示：

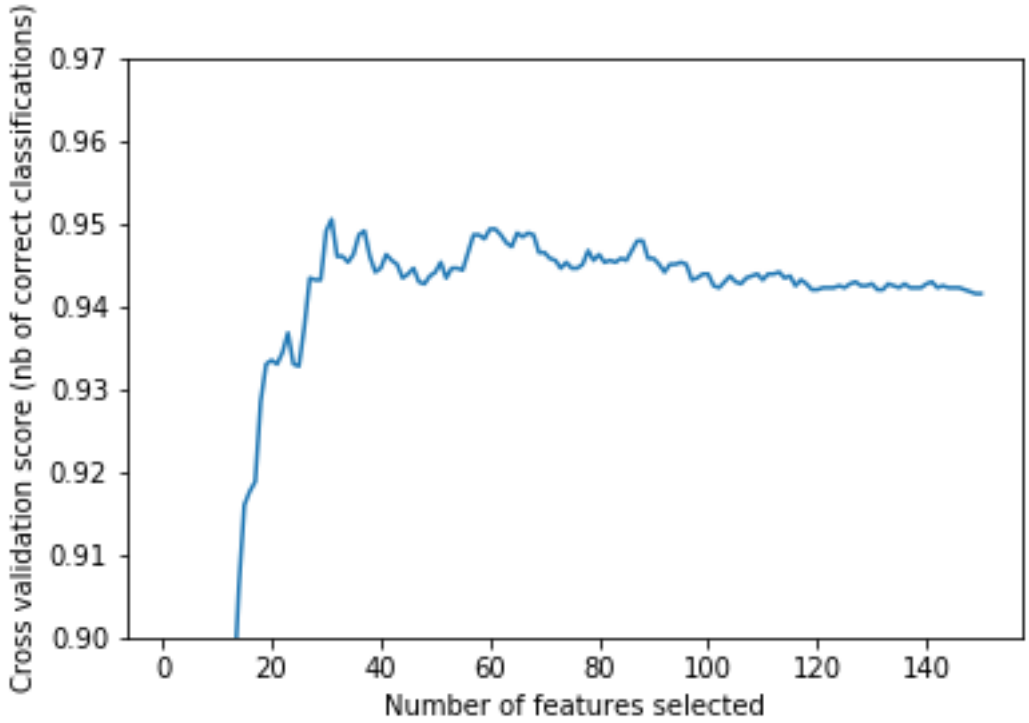

图 1 交叉验证分数（平均准确率）与特征集大小的对应关系

需要注意的是，识别准确率首次能够达到 80%以上的特征集大小是 7 维，此时识别准确率是 81.37%；识别准确率首次能够达到 90%以上的特征集大小是 13 维，此时识别准确率是 90.61%。这说明，在最佳维度数量之前，识别准确率随特征集大小增大而快速提高，但接近峰值，识别准确率随特征维度增加而只能缓慢提高，这意味着获得相同百分比的准确率提升需要付出更多的计算空间和运算时间。实际应用中往往会在运算效率和识别准确率之间做一个平衡。在最佳维度数量之后，识别准确率随特征数量增加在 94.5%附近小范围波动起伏，最低是 94.15%。这说明特征集内部存在冗余性，更多的特征不会对识别准确率有提升，反而会降低乐器识别系统的性能。

31 维最佳特征集如下表所示：

表格 4 最佳特征集

特征类别	特征中文名称	特征英文名称
感知特征	相关响度标准	loudness_ebu128.integrated
		loudness_ebu128.momentary.median
		loudness_ebu128.short_term.median
	不协和度	dissonance.median
	动态复杂度	dynamic_complexity

能量特征	总能量（Total Energy）	ERBfft_FrameErg_median
倒谱域特征	梅尔频率倒谱系数（MFCC）	mfcc_mean1
		mfcc_mean2
		mfcc_mean3
		mfcc_mean4
		mfcc_mean5
		mfcc_mean8
		mfcc_mean11
		mfcc_mean12
时域特征	时域质心（Temporal Centroid）	TEE_TempCent
频域统计特征	谱分布方差（Spectral Spread）	STFTmag_SpecSpread_median
	谱偏态（Skewness）	STFTpow_SpecSkew_median
	谱通量（Spectral Variation、Spectral Flux）	STFTmag_SpecVar_median
	谱下降值（Spectral Roll-off）	STFTpow_SpecRollOff_median
	谱分布方差（Spectral Spread）	STFTpow_SpecSpread_median
	谱峰值因子（Spectral Crest）	STFTpow_SpecCrest_median
	谱熵（Spectral Entropy）	spectral_entropy.median/stdev
谐波特征	噪声度（Noisiness）	Harmonic_Noisiness_median
听觉模型频谱包络统计特征	谱峰度-等效矩形带宽（ERB-Perceptual Spectral Kurtosis）	ERBfft_SpecKurt_median
	谱下降值-等效矩形带宽（ERB-Perceptual Spectral Roll-off）	ERBfft_SpecRollOff_median
	谱通量-等效矩形带宽（ERB-Perceptual Spectral Variation）	ERBfft_SpecVar_median
	谱平整度-等效矩形带宽（ERB-Perceptual	ERBfft_SpecFlat_median

	Spectral Flatness）	
	谱平整度-梅尔频带（Mel bands-Perceptual Spectral Flatness）	melbands_flatness_db.median/stdev
	谱平整度-巴克频带（Bark bands -Perceptual Spectral Flatness）	barkbands_flatness_db.median

通过观察最佳特征集的组成，可以得到以下的分析结论：

1、频谱信号统计特征占据主体（48.39%），这说明此类特征用于乐器识别的高效性。

2、倒谱域特征中梅尔倒谱频率系数（MFCC）的一阶差分、二阶差分均没有入选，这说明它们的重要性比不上未经差分处理的系数。而在 MFCC 的 13 个系数中，前 5 个全部入选，后 8 个只入选了 3 个，则反映出 MFCC 内部系数的重要性排序情况。

3、时域能量包络特征、能量特征和谐波特征分别只有一个特征被选中，而时域特征以及谐波频谱包络统计特征均没有被选中，可见对于中国民族识别来说，这些特征类别的重要性比不上最佳特征集内的特征。从信号表征的角度来分析，四种音频信号表征中，希尔伯特变换和正弦谐波模型分别只有一个入选，而其余两种表征短时傅里叶变换和听觉模型均有大量特征入选，而且占据最佳特征集的主体，可以认为，短时傅里叶变换和听觉模型这两种音频表征更适合应用于中国民族乐器分类问题中。

4、通过比较最佳特征集中的特征值统计方法，会发现中位数（Median）占据了主导地位（58.06%），而四分位距均没有入选，这说明了反应数据偏差程度的统计量在中国民族乐器识别中不太重要。

（五）最佳特征集乐器识别实验结果与分析

得到 31 维特征集后，开始进行乐器识别实验。首先使用训练集数据，对基于支持向量机的分类模型进行训练，支持向量机的内核函数为高斯函数(RBF)。需要注意的是，测试集数组也需要提前经过变换处理(Transform)，使其特征与最佳特征集保持一致，否则无法运行分类算法。实验所得到的结果如下表所示。

表格 5 交叉验证结果（最佳特征集+支持向量机）

乐器类别（按 F1-score 由低到高排序）	召回率 Recall	精确率 Precision	F1-score
Dacha 大镲	0.33	1	0.5
Xindi 新笛	0.55	0.73	0.63

Qudi 曲笛	0.84	0.68	0.75
Xiaotanggu 小堂鼓	0.67	1	0.8
Xiao 箫	0.8	0.94	0.86
Konghou 箜篌	0.86	1	0.92
Xun 埙	0.9	1	0.95
Bangdi 梆笛	0.97	0.97	0.97
Erhu 二胡	0.99	0.96	0.97
Alto Banhu 中音板胡	0.98	0.98	0.98
Guzheng 古筝	1	0.95	0.98
Pipa 琵琶	0.98	1	0.99
Sanxian 三弦	0.99	0.99	0.99
Soprano Banhu 高音板胡	0.98	1	0.99
Zhongruan 中阮	0.99	0.98	0.99
Chinese Dagu 中国大鼓	1	1	1
Hulusi 葫芦丝	1	1	1
LiujiaoGaohu 六角高胡	1	1	1
Liuqin 柳琴	1	0.99	1
Nao 铙	1	1	1
Naobo 铙钹	1	1	1
Sheng 传统笙	1	1	1
Suona 唢呐	1	1	1
Xiaocha 小镲	1	1	1
Yehu 椰胡	1	1	1
Yunluo 云锣	1	1	1
准确率 Accuracy	0.98		

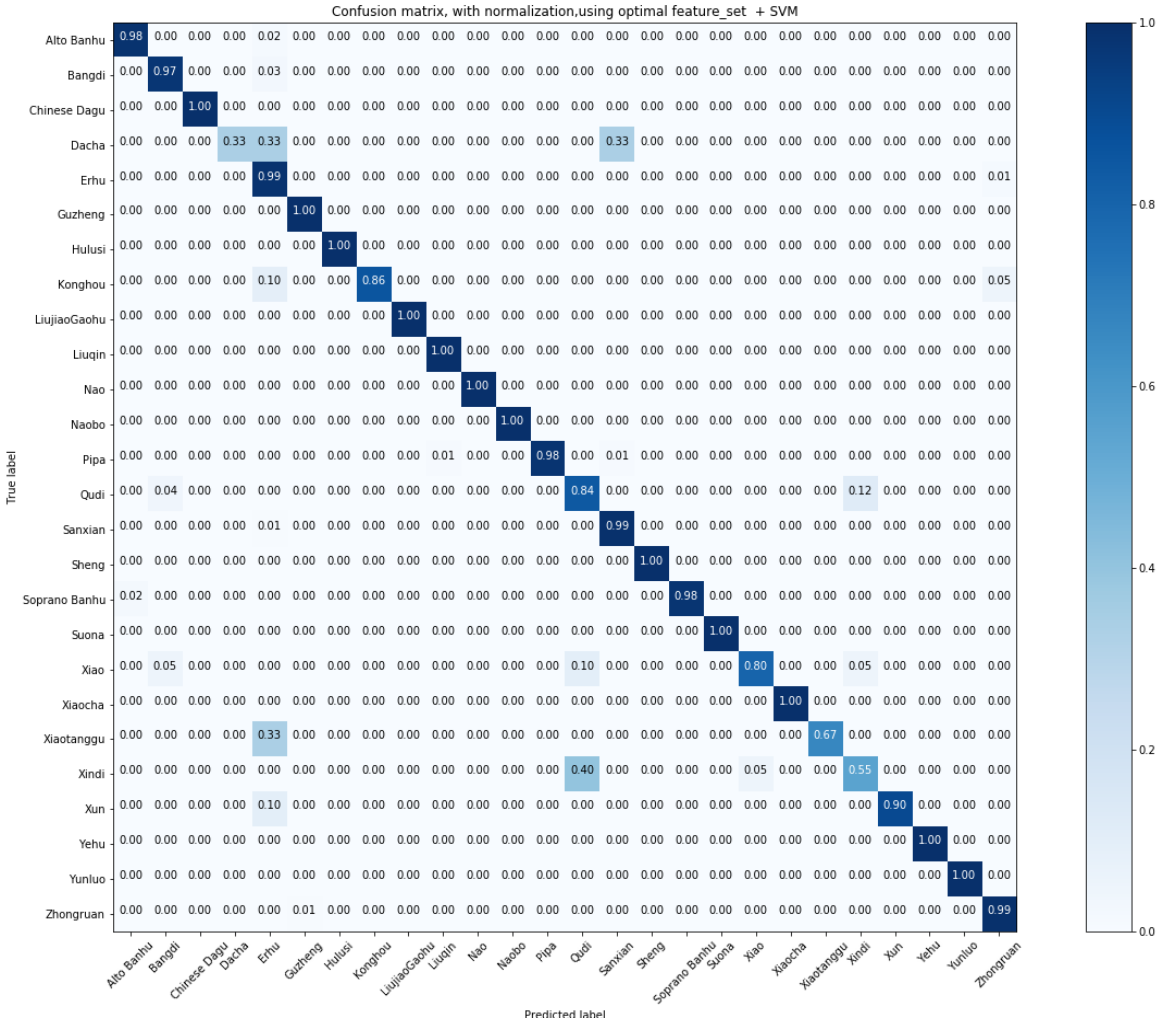

图 2 混淆矩阵（最佳特征集+支持向量机）

由表格 4 可知，由最佳特征集训练出来的分类模型，对于绝大部分乐器的分类情况十分不错,总体识别准确率为 98%。在 26 件乐器中，21 件乐器的 F1-score 超过 0.9，15 件乐器的 F1-score 非常接近 1。

F1-score 值较低的乐器有两种，一种是打击乐器（大镲、小堂鼓），另一种是吹管乐器（新笛、曲笛、箫）。前者的情况是因为训练样本的数量太少，另外除云锣外其他打击乐器样本中常规音色和特殊音色的比例失调，特殊音色的比例过高。后者的情况，是比较容易理解的，吹管乐器组内各件乐器音色上相似度较高。由混淆矩阵可以观察到，曲笛有 4%和 12%的样本分别与梆笛、新笛进行混淆，箫有 5%、5%和 10%的样本分别与梆笛、新笛以及曲笛进行混淆，新笛有 40%和 5%的样本分别于曲笛和箫进行混淆。

为了突出经过递归特征消除后乐器分类系统的改善，实验也对未经过递归特征消除的特征集进行模型训练。得到的结果如下：

表格 6 交叉验证结果（150 维特征集+支持向量机）

乐器类别（按 F1-score 由低到高排序）	召回率 Recall	精确率 Precision	F1-score
Dacha 大镲	0	0	0
Nao 铙	0	0	0
Naobo 铙钹	0	0	0
Xiaotanggu 小堂鼓	0	0	0
Xun 埙	0	0	0
Xindi 新笛	0.1	0.67	0.17
Yunluo 云锣	0.14	1	0.25
Xiao 箫	0.15	1	0.26
Qudi 曲笛	0.16	1	0.28
Chinese Dagu 中国大鼓	0.33	1	0.5
Konghou 箜篌	0.33	1	0.5
Xiaocha 小镲	0.33	1	0.5
Erhu 二胡	1	0.35	0.52
Yehu 椰胡	0.4	1	0.57
Hulusi 葫芦丝	0.49	1	0.66
LiujiaoGaohu 六角高胡	0.49	1	0.66
Bangdi 梆笛	0.5	1	0.67
Suona 唢呐	0.57	1	0.72
Soprano Banhu 高音板胡	0.59	1	0.74
Liuqin 柳琴	0.64	0.96	0.77
Alto Banhu 中音板胡	0.67	0.96	0.79
Guzheng 古筝	0.65	1	0.79
Sheng 传统笙	0.68	1	0.81
Sanxian 三弦	0.84	0.99	0.91
Pipa 琵琶	0.9	0.98	0.94
Zhongruan 中阮	0.91	0.99	0.95
准确率 Accuracy	0.71		

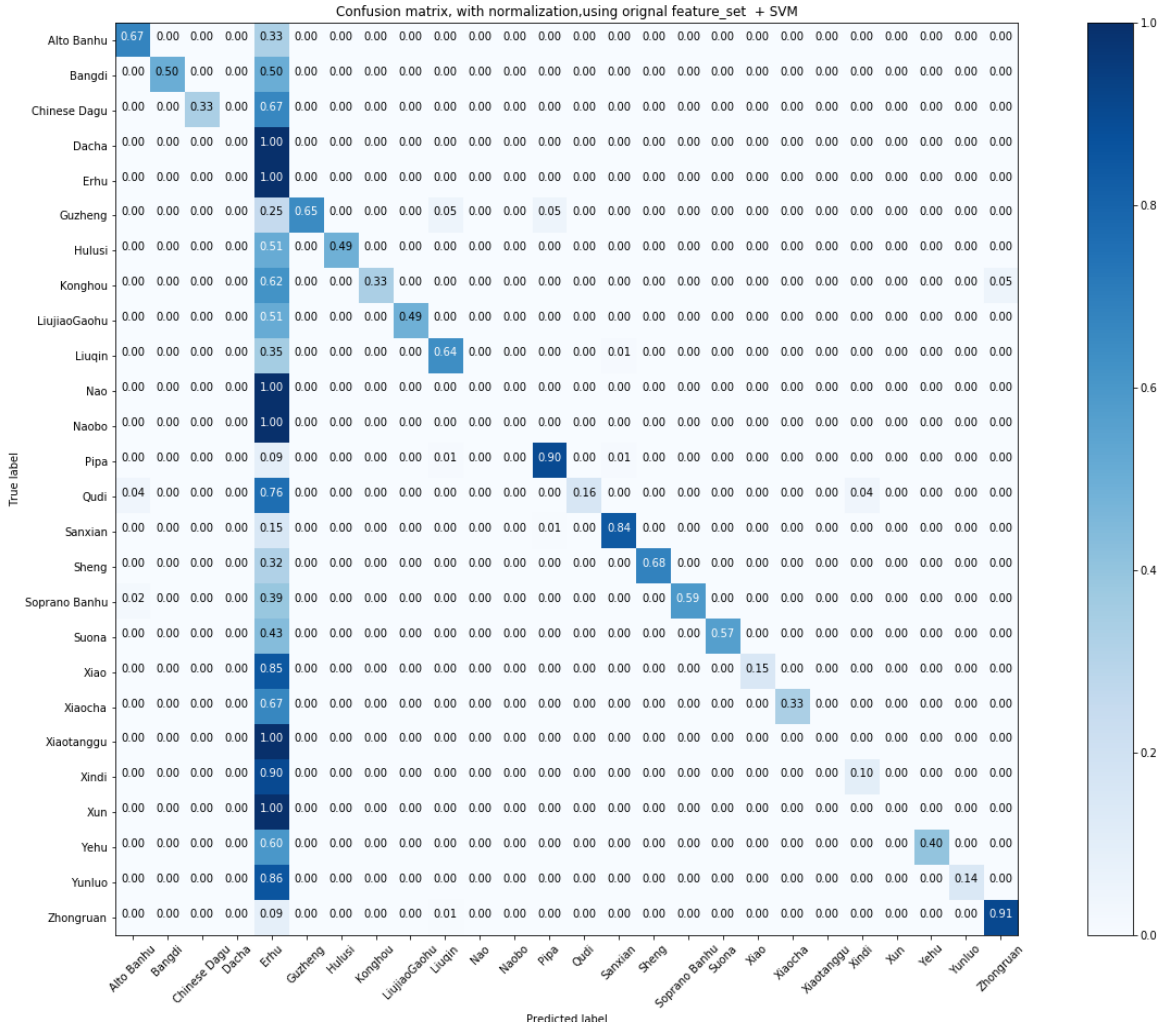

图 3 混淆矩阵（150 维特征集+支持向量机）

通过观察图 3 混淆矩阵，可以清晰地发现，很多乐器都被错误地分类成二胡，而不容易被错误分类的乐器有：中阮、琵琶、三弦、二胡。这种状况和该乐器的样本数有明显的相关性：样本数较多的乐器不容易被错分，而样本数较少的乐器则几乎全部错分成样本数最多的乐器二胡。这暴露了本实验数据库中各个乐器样本数量之间的不平衡。

由表格 6 可知，基于 150 维原始特征集训练出来的分类模型，对于所有测试集中的打击乐器都无法做到较好的分类，所有打击乐器的 F1-score 均低于 0.5。另外，由图 1 可知，对于训练集，该分类模型可以获得 94%左右的识别准确率，而现在对于测试集，分类模型只能获得 71%的识别准确率，对大部分乐器展现出低召回率和高精确率的情况，这说明分类模型存在高方差问题。其中一个原因就是所选用了过多特征，分类模型对训练数据出现了过拟合（Overfitting）的现象。因此，通过特征选择方法来适当减少分类模型的输入特征十分有必要，它有助于提高分类模型对未来测试样本的泛化能力。

五、结论

本论文的主要工作是：参考其他学者的特征分析研究，整理出一个比较全面的用于乐器单音识别的特征分类框架，并利用 Timbre Toolbox、Essentia 以及 Librosa，对 26 件中国民族乐器共 5654 个单乐音数据库构建出 235 维初始特征集。随后使用包裹式特征选择方法——递归特征消除（RFE），来得出最佳特征集的维度数、同时完成 31 维最佳特征集的构建。最后将最佳特征集用在基于支持向量机的分类模型训练上，并获得了 98%的平均准确率。对比没有经过特征选择的 150 维原始特征集所训练出来的分类模型，最佳特征集令分类模型获得了 27%准确率的提升，而且大大减轻了过拟合现象，这说明特征选择能够极大地优化分类系统的性能。

由于本科阶段知识结构的不完善和分析能力的有限，本实验还存在一些问题和进步的空间。首先是数据库中部分乐器的样本太少，如打击乐器常规音色的样本，这导致最佳特征集在大镲和小堂鼓的识别表现上不够优秀，这可能是 DCMI 数据库建立时候所疏忽的地方。其次，实验发现最佳特征集在吹奏乐器族内识别的表现不够优秀，未来可尝试手动添加特征，来实现对所有乐器的高效识别。最后，由于本实验的乐器分类系统只是基于支持向量机而建立起来，这会让本文所提出的 31 维最佳特征集具有一定局限性，未来可尝试整合多种分类算法，来比较最佳特征集在不同分类算法上的表现，以获取更可靠的实验结论。

参考文献

[1] J. D. Deng, C. Simmermacher and S. Cranefield, "A Study on Feature Analysis for Musical Instrument Classification," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 38, no. 2, pp. 429-438, April 2008, doi: 10.1109/TSMCB.2007.913394.

[2] Mingchun Liu and Chunru Wan. 2001. Feature selection for automatic classification of musical instrument sounds. In Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries (JCDL '01). Association for Computing Machinery, New York, NY, USA, 247–248. DOI:https://doi.org/10.1145/379437.379663

[3] P. Uruthiran and L. Ranathunga, "Optimization of Feature Selection and Classification of Oriental Music Instruments Identification," 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), Ipoh, Perak, Malaysia, 2019, pp. 120-125, doi: 10.1109/AiDAS47888.2019.8970974.

[4] 后方帅. 基于单音的西洋乐器音色识别方法研究[D]. 硕士学位论文，山东大学, 2018.

[5] J. Liu and L. Xie, "Comparison of Performance in Automatic Classification between Chinese and Western Musical Instruments," 2010 WASE International Conference on Information Engineering, Beidaihe, Hebei, 2010, pp. 3-6, doi: 10.1109/ICIE.2010.8.

[6] 刘璇. 基于多特征组合分类的乐器识别研究[C]. 中国计算机学会多媒体技术专业委员会 、中国图象图形学学会多媒体专业委员会 、中国计算机学会普适计算专业委员会、ACM SIGCHI 中国分会、中国自动化学会.第六届和谐人机环境联合学术会议（HHME2010)、第 19 届全国多媒体学术会议（NCMT2010）、第 6 届全国人机交互学术会议（CHCI2010）、第 5 届全国普适计算学术会议（PCC2010）论文集.中国计算机学会多媒体技术专业委员会 、中国图象图形学学会多媒体专业委员会 、中国计算机学会普适计算专业委员会、ACM SIGCHI 中国分会、中国自动化学会:中国计算机学会多媒体技术专业委员会,2010:148-153.

[7] 杨婧. 基于谐波结构的乐器音色特征提取方法研究[D]. 硕士学位论文，哈尔滨工业大学, 2018.

[8] 龙丽婧. 西洋乐器的音色特征值提取与研究[D]. 硕士学位论文，上海师范大学, 2011.

[9] 王琪. 西洋乐器的音色识别[D]. 硕士学位论文，山东大学, 2015.

[10] Bahre, Shubham & Mahajan, Shrinivas & Pillai, Rohan. (2017). "Novel audio feature set for monophonie musical instrument classification." 562-565. 10.1109/RISE.2017.8378218.

[11] G. S. R., B. S. S. and S. S. D., "Cepstral (MFCC) Feature and Spectral (Timbral) Features Analysis for Musical Instrument Sounds," 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN), Lonavala, India, 2018, pp. 109-113, doi: 10.1109/GCWCN.2018.8668628.

[12] 刘若伦,张家琦.乘法模型下西洋乐器音色特征[J].声学技术,2009,28(03):269-275.

[13] Volume 46 of the Zeitschrift , pages 553–90, 1914

[14] Li, Zijin and Xiaojing Liang. “DCMI: A Database of Chinese Musical Instruments.” (2018).

[15] Grey, J., and Gordon, J.. “Perceptual effects of spectral modifications on musical timbres,” J. Acoust. Soc. Am. 1978, 63, 1493-1500.

[16] W. Jiang, J. Liu, Z. Li, J. Zhu, X. Zhang and S. Wang, "Analysis and Modeling of Timbre Perception Features of Chinese Musical Instruments," 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 2019, pp. 191-195, doi: 10.1109/ICIS46139.2019.8940168.

[17] Peeters Geoffroy,Giordano Bruno L,Susini Patrick,Misdariis Nicolas,McAdams Stephen. The Timbre Toolbox: extracting audio descriptors from musical signals.[J]. Pubmed,2011,130(5)

[18] Serra, X., and Smith J. “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” Comput. Music J. 1990, 14, 12-24.

[19] R. Plomp and W. J. M. Levelt, "Tonal Consonance and Critical Bandwidth," The Journal of the Acoustical Society of America, vol. 38, no. 4, pp. 548–560, 1965.

附录

表格 7 原始特征集（完整）

特征类别	特征名称	在提取工具中的命名	维数
能量特征	总能量（Total Energy）	STFTmag/pow_FrameErg_median/iqr ERBfft/gam_FrameErg_median/iqr	8
	总能量调制（Total Energy Modulation）	TEE_AmpMod	1
	均方根能量（RMS Energy）	TEE_RMSEnv_median/iqr	2
	正弦谐波模型总能量（Total Frame Energy）	Harmonic_FrameErg_median/iqr	2
	总谐波能量（Harmonic Energy）	Harmonic_HarmErg_median/iqr	2
	总噪声能量（Noise Energy）	Harmonic_NoiseErg_median/iqr	2
时域包络特征	对数起振时间（Log Attack Time）	TEE_LAT	1
	时域起振（Temporal Increase）	TEE_Att、TEE_AttSlope	2
	时域衰减（Temporal Decrease）	TEE_Dec、TEE_Rel、TEE_DecSlope	3
	时域质心（Temporal Centroid）	TEE_TempCent	1
	有效时长（Effective Duration）	TEE_EffDur	1
时域特征	信号自相关函数（Signal Auto-Correlation Function）	AS_AutoCorr_median/iqr	24
	过零率（Zero Crossing Rate）	AS_ZcrRate_median/iqr	2
频域频谱包络统计特征	谱质心（Spectral Centroid）	STFTmag/pow_SpecCent_median/iqr	4
	谱分布方差（Spectral Spread）	STFTmag/pow_SpecSpread_median/iqr	4
	谱偏态（Skewness）	STFTmag/pow_SpecSkew_median/iqr	4
	谱峰度（Spectral	STFTmag/pow_SpecKurt_median/iqr	4

	Kurtosis）		
	谱斜度（Spectral Slope）	STFTmag/pow_SpecSlope_median/iqr	4
	谱衰减（Spectral Decrease）	STFTmag/pow_SpecDecr_median/iqr	4
	谱下降值（Spectral Roll-off）	STFTmag/pow_SpecRollOff_median/iqr	4
	谱通量（Spectral Variation、Spectral Flux）	STFTmag/pow_SpecVar_median/iqr	4
	谱平整度（Spectral Flatness）	STFTmag/pow_SpecFlat_median/iqr	4
	谱峰值因子（Spectral Crest）	STFTmag/pow_SpecCrest_median/iqr	4
	谱熵（Spectral Entropy）	lowlevel.spectral_entropy	2
谐波特征	基频（Fundamental Frequency）	Harmonic_F0_median/iqr	2
	基频调制（Fundamental Frequency Modulation）	TEE_FreqMod	1
	噪声度（Noisiness）	Harmonic_Noisiness_median/iqr	2
	不和谐度（Inharmonicity）	Harmonic_InHarm_median/iqr	2
	谐波谱偏差（Harmonic Spectral Deviation）	Harmonic_HarmDev_median/iqr	2
	奇偶谐波比例（Odd-to-Even Harmonic Ratio）	Harmonic_OddEvenRatio_median/iqr	2
	谐波三刺激值（Harmonic Tristimulus）	Harmonic_TriStim（1、2、3）_median/iqr	6
谐波频谱包络统计特征	谐波谱质心（Harmonic Spectral Centroid）	Harmonic_SpecCent_median/iqr	2
	谐波谱分布方差（Harmonic Spectral Spread）	Harmonic_SpecSpread_median/iqr	2
	谐波谱偏态（Harmonic Skewness）	Harmonic_SpecSkew_median/iqr	2
	谐波谱峰度	Harmonic_SpecKurt_median/iqr	2

	（Harmonic Spectral Kurtosis）		
	谐波谱斜度（Harmonic Spectral Slope）	Harmonic_SpecSlope_median/iqr	2
	谐波谱衰减（Harmonic Spectral Decrease）	Harmonic_SpecDecr_median/iqr	2
	谐波谱下降值（Harmonic Spectral Roll-off）	Harmonic_SpecRollOff_median/iqr	2
	谐波谱通量（Harmonic Spectral Variation）	Harmonic_SpecVar_median/iqr	2
感知特征	平均响度（Loudness）	lowlevel.average_loudness	1
	相关标准响度（Relative Specific Loudness）	lowlevel.loudness_ebu128.integrated、lowlevel.loudness_ebu128.loudness_range、lowlevel.loudness_ebu128.momentary lowlevel.loudness_ebu128.short_term	6
	不协和度（Dissonance）	lowlevel.dissonance	2
	动态复杂度（Dynamic Complexity）	lowlevel.dynamic_complexity	1
听觉模型频谱包络统计特征	谱质心-等效矩形带宽（ERB-Perceptual Spectral Centroid）	ERBfft/gam_SpecCent_median/iqr	4
	谱分布方差-等效矩形带宽（ERB-Perceptual Spectral Spread）	ERBfft/gam_SpecSpread_median/iqr	4
	谱偏态-等效矩形带宽（ERB-Perceptual Spectral Skewness）	ERBfft/gam_SpecSkew_median/iqr	4
	谱峰度-等效矩形带宽（ERB-Perceptual Spectral Kurtosis）	ERBfft/gam_SpecKurt_median/iqr	4
	谱斜度-等效矩形带宽（ERB-Perceptual Spectral Slope）	ERBfft/gam_SpecSlope_median/iqr	4
	谱下降-等效矩形带宽（ERB-Perceptual Spectral Decrease）	ERBfft/gam_SpecDecr_median/iqr	4
	谱下降值-等效矩形带	ERBfft/gam_SpecRollOff_median/iqr	4

	宽（ERB-Perceptual Spectral Roll-off）		
	谱通量-等效矩形带宽（ERB-Perceptual Spectral Variation）	ERBfft/gam_SpecVar_median/iqr	4
	谱平整度-等效矩形带宽（ERB-Perceptual Spectral Flatness）	ERBfft/gam_SpecFlat_median/iqr	4
	谱峰值因子-等效矩形带宽（ERB-Perceptual Spectral Crest）	ERBfft/gam_SpecCrest_median/iqr	4
	谱分布方差-梅尔频带（Mel bands-Perceptual Spectral Spread）	lowlevel.melbands_spread	2
	谱偏态-梅尔频带（Mel bands - Perceptual Spectral Skewness）	lowlevel.melbands_skewness	2
	谱峰度-梅尔频带（Mel bands - Perceptual Spectral Kurtosis）	lowlevel.melbands_kurtosis	2
	谱峰值因子-梅尔频带（Mel bands - Perceptual Spectral Crest）	lowlevel.melbands_crest	2
	谱平整度-梅尔频带（Mel bands - Perceptual Spectral Flatness）	lowlevel.melbands_flatness_db	2
	谱分布方差-巴克频带（Bark bands - Perceptual Spectral Spread）	lowlevel.barkbands_spread.mean	2
	谱偏态-巴克频带（Bark bands - Perceptual Spectral Skewness）	lowlevel.barkbands_skewness.mean	2
	谱峰度-巴克频带（Bark bands - Perceptual Spectral	lowlevel.barkbands_kurtosis.mean	2

	Kurtosis）		
	谱峰值因子-巴克频带（Bark bands - Perceptual Spectral Crest）	lowlevel.barkbands_crest.mean	2
	谱平整度-巴克频带（Bark bands - Perceptual Spectral Flatness）	lowlevel.barkbands_flatness_db.mean	2
倒谱域特征	梅尔频率倒谱系数（MFCC）	librosa.feature.mfcc	13
	梅尔频率倒谱系数一阶差分（DMFCC）	librosa.feature. delta（mfcc）	13
	梅尔频率倒谱系数二阶差分（DDMFCC）	librosa.feature. delta（mfcc, order=2）	13
总维数			235

表格 8 Timbre Toolbox 提取谐波特征的错误情况

乐器类别（按错误率由高到低排序）	无法检测谐波特征的样本数（个）	总样本数量（个）	所占百分比
铙钹 Naobo	16	16	100.00 %
小镲 Xiaocha	12	12	100.00 %
高音板胡 Soprano Banhu	164	197	83.25 %
唢呐 Suona	157	213	73.71 %
铙 Nao	11	15	73.33 %
六角高胡 LiujiaoGaohu	137	187	73.26 %
中音板胡 Alto Banhu	161	256	62.89 %
G 调梆笛 Bangdi	200	321	62.31 %
大镲 Dacha	8	14	57.14 %
传统笙 Sheng	138	250	55.20 %
云锣 Yunluo	31	58	53.45 %
中国大鼓 Chinese Dagu	9	23	39.13 %
二胡 Erhu	237	870	27.24 %
A 调曲笛 Qudi	25	99	25.25 %
琵琶 Pipa	111	495	22.42 %
古筝 Guzheng	16	80	20.00 %
椰胡 Yehu	16	80	20.00 %

箜篌 Konghou	15	86	17.44 %
小堂鼓 Xiaotanggu	2	12	16.67 %
G 调新笛 Xindi	12	82	14.63 %
三弦 Sanxian	95	750	12.67 %
柳琴 Liuqin	53	495	10.71 %
箫 Xiao	5	79	6.33 %
埙 Xun	2	41	4.88 %
中阮 Alto Ruan	33	720	4.58 %
葫芦丝 Hulusi	4	203	1.97 %
总数统计	1670	5654	29.54 %

后记

非常感谢父母、学院老师、同学和朋友给予我的支持和鼓励。是你们的爱让我充满力量、奋勇向前、力争上游。我时常感受到自己被爱包围，我是何其幸运。

经历了整个毕业设计过程，我在很多方面得到了提升。其中很重要的一点，是发现问题、解决问题的能力。实验过程中遇到不少问题，比如说在使用 Matlab 提取特征的时候，“六角高胡”、“唢呐”、“大镲”、“G 调梆笛（特殊技巧部分音频）”这四个种类的样本集中出现报错，Matlab 显示“矩阵维度不匹配”。寻找问题源头，发现 DCMI 数据库中上述乐器音频文件的采样频率是 48kHz，与其余音频文件的采样频率 44.1kHz 不同，这导致矩阵维度不匹配。我的解决方法是，使用 Adobe Audition CC 对上述采样频率不相符的音频文件进行采样频率转换。这也提醒了我要对数据库中的音频做好检查和统一参数的工作。

另一个问题是：存在部分样本，由于它们处于系统参数谐波阈值（config_s.threshold_harmo）之下，Timbre Toolbox 没有对其进行谐波特征提取。这类样本共 1670 个，占样本总数 29.5%。各乐器中错误样本的详细比例可见附录表格 8。由表格 8 可知，错误率较高的样本组成主要是打击乐器样本以及基频高的样本（高音乐器、普通乐器高音的乐音、以及部分特殊技巧如泛音）。其中打击乐器有超过 70%的乐器特征提取错误率超过 50%。曾猜测 Timbre Toolbox 设置谐波阈值背后的指导思想，可能是希望对打击乐器不使用正弦谐波模型（Sinusoidal Harmonic Model）进行研究，因为此类乐器本身是一种噪音乐器，正弦谐波在乐音中的比例不高。对于其他中国民族乐器的特征提取，Timbre Toolbox 的默认参数不太适用于高音乐器，如高音板胡、六角高胡、唢呐、中音板胡。通过降低谐波阈值，可以实现对所有乐器的谐波特征提取，但发现谐波特征中的基频 F0 基本都不准确（比实际基频低了一个八度），这可能是 Timbre Toolbox 内部算法的设置问题。由于不肯定擅自调低谐波阈值是否对于其他谐波特征的提取产生影响，本实验中还是按照默认参数进行特征提取。因此，由于有近三成的样本没有进行谐波特征，这可能是最佳特征集中缺少谐波频谱包络统计特征的原因。如果将来有同行使用 Timbre Toolbox 对中国民族乐器的高音乐器进行特征提取，希望可以注意到这个问题的存在。

最后，谨以此文献给身边给予我关爱的人。