基于降维聚类与强化学习的高相似度弦乐器音色区分及合成优化研究

赵星媛[1] 白 羽[1] 沈函宇[1] 唐惠心[1] 魏依琳[1] 宋 飞[1,2] 张留碗[2]

([1]清华大学行健书院,北京 100084;[2]清华大学物理系,北京 100084)

摘 要 针对音频合成音色自然度不足的问题,本研究提出一种融合特征分析与智能优化的多阶段框架。首先,通过合成三类声波构建数据集,提取声学特征,利用降维与聚类量化合成音与自然音的差异;其次,基于贝叶斯优化自适应分配特征权重,筛选关键区分性指标;最后,结合强化学习动态调整合成参数,以聚类中心距离为奖励驱动音色逼近自然声分布。实验证明,本方法显著提升合成音色的自然度,为数据驱动的音频优化提供了高效解决方案。

关键词 音频合成;声学特征;强化学习;贝叶斯优化;音色优化

A STUDY ON DIFFERENTIATION AND SYNTHESIS OPTIMIZATION OF HIGHLY SIMILAR STRING INSTRUMENT TIMBRES BASED ON DIMENSIONALITY REDUCTION, CLUSTERING, AND REINFORCEMENT LEARNING

ZHAO Xingyuan[1] BAI Yu[1] SHEN Hanyu[1] TANG Huixin[1] WEI Yilin[1] SONG Fei[1,2] ZHANG Liuwan[2]

([1] Xingjian College, Tsinghua University, Beijing 100084; [2] Department of Physics, Tsinghua University, Beijing 100084)

Abstract To address the insufficient naturalness of timbre in audio synthesis, this study proposes a multi-stage framework integrating feature analysis and intelligent optimization. First, a dataset is constructed by synthesizing three types of sound waves, from which acoustic features are extracted. Dimensionality reduction and clustering are employed to quantify the differences between synthetic and natural sounds. Second, Bayesian optimization is applied to adaptively assign feature weights, identifying key discriminative indicators. Finally, reinforcement learning dynamically adjusts synthesis parameters (e.g., frequency, decay factor) using clustering center distance as a reward to drive synthetic timbre closer to natural sound distributions. Experimental results demonstrate that this method significantly enhances the naturalness of synthesized timbre, providing an efficient data-driven solution for audio optimization.

Key words audio synthesis; acoustic features; reinforcement learning; Bayesian optimization; timbre optimization

收稿日期:2025-06-15;**修回日期**:2025-06-30

基金项目:通专教育背景下大学实验物理教学内容的重构探索与实践、教育部拔尖计划2.0研究课题(20211007);无锡市“未来技术太湖创新基金”;清华大学本科教育教学改革经费(53410001325)资助。

通信作者:宋飞,songfei_thu@tsinghua.edu.cn。

引文格式:赵星媛,白羽,沈函宇,等. 基于降维聚类与强化学习的高相似度弦乐器音色区分及合成优化研究[J]. 物理与工程,2025,35(5):329-336.

Cite this article: ZHAO X Y, BAI Y, SHEN H Y, et al. A study on differentiation and synthesis optimization of highly similar string instrument timbres based on dimensionality reduction, clustering, and reinforcement learning[J]. Physics and Engineering, 2025, 35(5): 329-336. (in Chinese)

随着数字音频技术的快速发展，音频合成在音乐制作、虚拟现实、语音合成等领域展现出广泛应用。然而，传统合成方法生成的音色往往在表现自然乐器的复杂声学特性上有所欠缺，导致合成音与自然音在听觉感知上存在显著差异。这种差异主要源于合成音在频谱结构、泛音分布及动态衰减特性上的简化建模。尽管物理建模算法能够模拟乐器振动特性，但其参数调优仍依赖经验，缺乏基于数据驱动的系统化优化框架。此外，现有研究对音色可分辨性的定量分析不足，尤其在声学特征权重分配与参数优化方面尚未形成普适性方法。因此，如何通过声学特征分析与机器学习技术，实现合成音色的智能化优化，成为提升音频合成自然度的关键挑战。

本研究拟通过三阶段递进式框架实现音色合成的智能化优化。首先，在数据构建与特征分析阶段，系统合成三类声波样本，并从中提取多维度声学特征。其次，借助 UMAP 非线性降维技术与 K-means 聚类算法，量化合成音与自然音在声学特征空间中的分布差异，通过调整兰德指数(ARI)客观评估聚类结果的一致性，为后续优化提供数据支撑。在特征权重优化阶段，以引入贝叶斯优化算法对多类声学特征进行自适应权重分配，通过迭代搜索确定对音色区分度影响最为显著的关键特征，为参数调优提供优先级依据。最后，在强化学习驱动合成优化阶段，设计基于策略网络的强化学习模型，以合成音与自然音聚类中心的距离作为动态奖励函数，实时调整频率、衰减因子等核心参数，逐步引导合成音在降维特征空间中逼近自然音的声学分布，从而实现音色自然度的显著提升。

本研究通过降维聚类算法与贝叶斯特征赋权的协同优化，实现了三种高相似度弦乐器：柳琴、琵琶、阮的音色的有效区分；设计优化模型并构建多指标评估框架完全区分人为调参的合成音。研究为音色自然度提升提供了数据驱动的优化路径，也为跨模态声学特征分析奠定了一定方法基础，具有较为显著的学术价值与广泛的应用潜力。

1 样本获取

1.1 试验设备

录音设备：48kHz 音频采样率，灵敏度－35dB，信噪比－65dB 麦克风，如图 1 所示。

图 1 自然音频录制设备

1.2 自然音频样本采集

使用琵琶、中阮、柳琴三种弦乐器，以近似相同拨弦力度，统一采集四根琴弦 1/2 与下 1/3 弦长处实按音与自然泛音音频[4]。受限于三种乐器定弦与音域差异，并排除不同位置同音高获得的重复音频，音频采集列表如表 1 所示。

表 1 自然音频采集列表

音名	d1	d2	d3	g	g1	g2	g3	a1	a2	a3
频率/Hz	294	587	1175	196	392	784	1568	440	880	1760
琵琶	√	√	×	×	s	√	f	√	√	×
柳琴	×	√	√	×	√	√	×	×	√	√
中阮	√	√	×	√	√	×	×	f	f	×

注：“√”代表实按音与泛音均有采集，“×”代表均未采集，“s”“f”分别代表仅采集实按音或泛音。

1.3 合成音频样本获取

正弦波合成得到稳定的基础单音[1]。基本公式

$$y(t)=A\sin(2f\pi t)$$

泛音合成通过叠加多个频率不同的正弦波来创建更复杂的音色[2]。调整不同谐波的振幅比例(生成随机权重)，确保各谐波振幅总和为 1，能够产生多样化的音色效果。基本公式为

$$y(t)=\sum_{i=1}^{n}A_i\sin(2f_i\pi t)$$

Karplus-Strong 算法利用延迟反馈和低通滤波的原理[3]，专门用于模拟合成弦乐器拨弦声音。Karplus-Strong 算法的核心原理是使用一个初始

的随机噪声缓冲区，模拟拨弦瞬间的扰动，然后通过一个带衰减的循环延迟线来重复播放这个扰动信号，同时在每次循环中来模拟能量的损失和高频的衰减，从而实现音色自然衰减和谐波结构的演化，最终产生类似真实拨弦乐器的声音。算法的核心公式为

$$y(n)=\alpha\frac{2}{y(n-1)+y(n-N)}$$

本次实验中，利用程序各生成 10 个不同频率和振幅的正弦波、泛音合成波和 Karplus-Strong 合成波，进行系统性的声音实验和比较分析。

2 方法论

本研究方法体系由声学特征处理、特征权重优化与人工声波合成三阶段构成。首先，提取三类合成音与三类自然音的声学特征，构建高维特征向量。其次，为探究特征空间的音色区分机制，对高维数据进行非线性降维，并通过聚类算法获得合成音与自然音的类别划分。引入 ARI 量化聚类标签与真实类别的一致性。再次，通过贝叶斯优化动态调整声学参数的权重，搜索使合成音与自然音区分度最大化的声学特征参数组合。最后，基于强化学习框架，通过奖励机制动态优化 Karplus-Strong 算法的合成参数，驱动合成音色逼近自然声的声学聚类中心。

2.1 声学特征提取

本研究采用 librosa 库（基于 Python 的音频和音乐分析库，广泛用于音频信号处理与机器学习任务中）从三类合成音与三类自然音中系统性提取声学特征，构建高维特征向量以支持后续分析。

首先，提取了 MFCC、色度特征、光谱特征和谐波与噪声比等参数，全面量化音色的物理属性[5,6]。MFCC（梅尔频率倒谱系数）通过模拟人耳听觉系统提取频谱包络，能够有效捕捉音色特征，常用于语音识别与音频分类。色度特征（Chroma）则将频谱映射到 12 个音高类，强调音乐中的调性和和声结构，适合用于旋律和和弦分析。光谱特征包括谱质心、谱带宽、谱对比度等，描述频谱的能量分布和明亮度等感知属性，反映音色细节。而 HNR（谐波与噪声比）则衡量声音中谐波与噪声的比例，揭示信号的周期性和清晰度。这些特征结合使用，可从频率结构、感知音色、调性信息和信号稳定性等多角度全面描述音频，为音乐信息检索、声音分类、音质评价等任务提供坚实基础。为增强特征的表征能力，对每类特征进一步计算统计量（均值、标准差、最大值、最小值、偏度及峰度），形成 217 维统计特征向量。这一多维特征组合不仅保留了音色的物理意义，还通过统计增强提升了特征的区分度，为后续降维、聚类及贝叶斯优化提供了高信息密度的数据基础。

2.2 降维与聚类

2.2.1 UMAP 降维算法

UMAP（Uniform Manifold Approximation and Projection）是一种基于流形学习的非线性降维算法，旨在将高维数据映射到低维空间的同时，尽可能保留数据的局部与全局结构。其核心原理

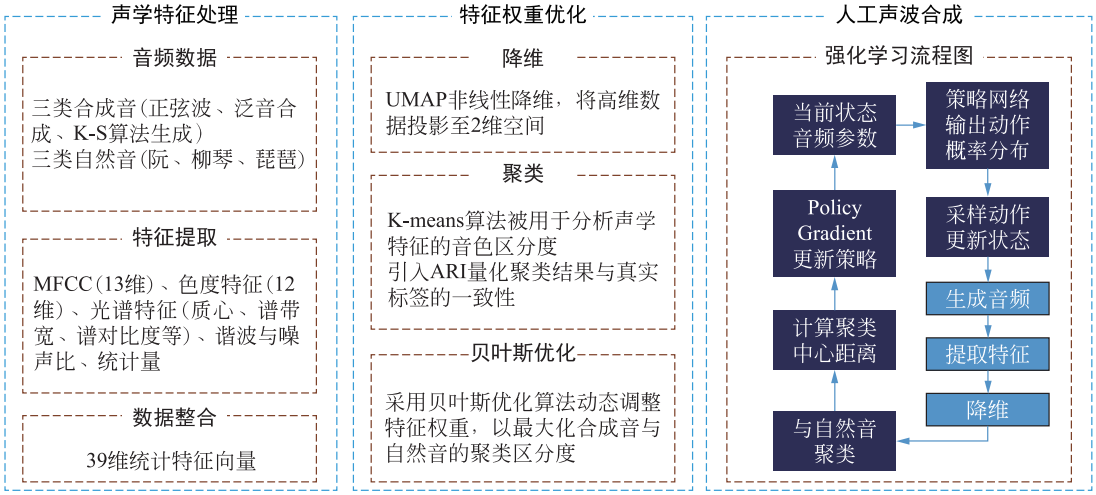

图 2 研究方法框架流程图

是通过构建高维数据的模糊拓扑表示模拟数据在流形上的分布，并利用交叉熵损失函数优化低维嵌入，使低维空间中的点间距离与高维拓扑关系相匹配。相较于传统方法，UMAP 在计算效率上显著提升，能够有效捕捉声学特征间的非线性关联、更精准量化音色区分度，且参数调节更直观[7]。

UMAP 的工作流程可以概括为两个主要阶段：第一，学习高维空间中的流形结构；第二，构建该流形在低维空间中的表示。具体而言，其原理流程如下所示：

1）高维空间的流形建模

UMAP 首先基于邻域搜索（通常采用近似最近邻方法，如 NNDescent）构建每个数据点的局部邻域，然后利用模糊集合理论将这些局部邻域表示为概率图，即以某种方式度量每对样本点属于同一局部结构的概率。这一过程实际上构造了一个加权无向图，用以描述高维空间中数据的局部拓扑关系。为了增强全局结构的信息保留，UMAP 还在构图过程中引入了平衡局部与全局的超参数，以控制图结构的“覆盖范围”。

2）低维空间的拓扑保持与优化嵌入

在构建完高维的模糊图之后，UMAP 在低维空间中初始化一个随机分布的点集，并尝试构造一个同样结构的低维图。随后，通过最小化高维图与低维图之间的交叉熵损失函数，不断调整低维嵌入中各点的位置，从而使低维空间中点与点之间的距离尽可能反映其在高维空间中所处的拓扑关系。这个优化过程既保留了局部邻域的紧密关系，也对较远的数据对进行了适度保留，从而在较低维度下同时体现数据的局部与全局结构。

2.2.2 K-means 算法原理与优势

K-means 算法作为一种经典的无监督聚类方法，在本研究中被用于分析降维后声学特征的音色区分度[8]。其核心目标是通过迭代优化将数据划分为预设的簇，使簇内样本相似性最大化。具体实现时，算法首先随机初始化两类质心，随后交替执行“样本分配”（基于欧氏距离将每个点归类至最近质心）与“质心更新”（重新计算簇内均值）步骤，直至质心位置收敛。为提升稳定性，本研究采用 K-means++策略优化初始质心选择，避免陷入局部最优解，依据样本情况调整迭代次数以确保计算效率与结果的可靠性。

选择 K-means 算法主要基于其高效性、可解释性与项目需求的适配性。一方面，其线性时间复杂度可高效处理本项目的音频样本数据集；另一方面，算法输出的簇质心能够直观表征合成音与自然音在低维空间中的“声学中心”，为强化学习的奖励函数设计提供直接依据。

2.2.3 聚类效果评估

在聚类效果评估中，本研究引入调整兰德指数（Adjusted Rand Index，ARI）量化聚类结果与真实标签的一致性。ARI 通过校正随机聚类的影响，将结果映射到[−1,1]区间，其中 1 表示完全匹配，0 为随机水平，负值则表明分类效果差于随机。ARI 的具体公式如下

$$\mathrm{ARI}=\frac{\mathrm{RI}_{\max}-E(\mathrm{RI})}{\mathrm{RI}-E(\mathrm{RI})}$$

其中，RI（Rand Index）衡量样本对在聚类与真实标签中的匹配比例[9]。

2.3 基于贝叶斯优化的特征权重分配

为量化声学特征对音色区分度的贡献，本研究采用贝叶斯优化算法动态调整特征权重，以最大化合成音与自然音的聚类区分度[10]。贝叶斯优化是一种基于概率模型的全局优化方法，通过构建目标函数的先验分布（通常为高斯过程），在每一步选择最可能带来改进的参数点进行评估，从而在尽量少的尝试下找到最优解。算法以 ARI 为评估基准，构建损失函数 $L=1-\mathrm{ARI}$，目标是通过最小化 L 提升聚类结果与真实类别的一致性。参数空间涵盖四类声学特征（MFCC、色度、光谱、谐波比）的权重 ω_i，搜索范围设定为[0.5,5.5]，通过缩放权重调节不同特征的影响力。

优化过程采用高斯过程作为代理模型，结合期望改进采集函数，平衡全局探索与局部开发。在 500 轮迭代中，每轮试验生成一组权重组合，加权后的特征矩阵经 UMAP 降维及 K-means 聚类（$k=2$）后计算 ARI 值，并更新代理模型以指导下一轮参数采样。当损失函数变化率低于阈值时终止优化，输出最优权重 ω 及最小损失值 L。通过分析权重分布（如 MFCC 权重显著高于色度特征），可解析不同声学属性对音色区分度的贡献，例如高频谐波比的权重提升可能关联音色明亮度的区分能力。

该方法优势在于其高效性与可解释性：贝叶斯优化通过历史评估结果引导搜索方向，避免网

格搜索的计算冗余；最优权重直接反映特征的物理意义，为强化学习的参数调整提供数据驱动依据。最终，优化权重被整合至后续流程，驱动合成音色逼近自然音的声学特性，凸显了贝叶斯优化在复杂特征空间中的全局优化能力。

2.4 基于声学特征区分度的强化学习声波合成方法

为实现合成音色与自然音色的逼近，本研究设计了基于策略梯度的强化学习框架，通过动态调整 Karplus-Strong 算法的核心参数（频率、振幅、衰减因子、持续时间），优化合成音色的声学特性[3]。

强化学习（Reinforcement Learning，RL）是一类旨在通过智能体与环境之间的动态交互，学习最优决策策略的机器学习方法。该框架通常形式化为一个马尔可夫决策过程，包括状态空间、动作空间、状态转移概率、奖励函数以及折扣因子。

在强化学习中，奖励函数是系统设计的关键组成部分，用于量化智能体在特定状态下采取特定动作所获得的即时反馈。该函数直接决定了智能体对环境反馈的敏感性与学习目标，是策略优化的核心驱动。

策略梯度方法是强化学习中一类直接对策略函数进行优化的经典方法。与基于值函数的方法不同，策略梯度方法显式对参数化策略建模，并通过最大化期望累计回报来学习策略参数。

在本模型中，策略网络采用全连接神经网络结构，输入归一化后的合成参数，输出 8 类离散动作的概率分布，模拟人工调参的探索与决策过程[11,12]。动作空间的设计与经典离散控制问题一致，确保参数调整的精细化和可解释性。

奖励函数以合成音与自然音在 UMAP 降维空间中的聚类质心距离为核心优化目标，定义为

$$R = - \| C_i - C_{\text{natural}} \|^2$$

通过 500 轮迭代训练，策略网络逐步学习最大化奖励的参数调整模式。每一轮训练中，新生成的合成音频与固定自然音样本共同提取声学特征，经统一模型降维后，输入 K-means 聚类（$k=2$），实时计算聚类中心距离并反馈至奖励机制。

最终，优化后的合成音与自然音在声学特征空间中的 ARI 趋近于随机水平，表明经优化后降维聚类算法难以区分两类音色，验证了合成效果的真实性。

3 实验结果讨论

本研究通过降维聚类与贝叶斯特征赋权的协同优化，实现了三种高相似度弦乐器：柳琴、琵琶、阮的音色的有效区分。实验表明，UMAP 在初步降维中虽能呈现样本聚集趋势，但三类乐器存在显著重叠，凸显传统方法的局限性。引入分块赋权策略后，动态优化 MFCC、色度特征、光谱特征及谐波与噪声比的权重，赋权后 ARI 提升至较优化前增加了 84.5%，类间距缩短，进一步证明光谱特征与 MFCC 参数为音色核心区分维度。合成音验证显示，优化模型在本实验数据集中可以近乎完全区分人为调参的合成音，而难以区分通过强化学习调参的合成音。本研究通过多指标评估框架：ARI、声学中心间距与可解释性赋权设计，为音频分类与合成提供了新范式。

3.1 数据采集与特征提取

实验采集了三种乐器在相同频率与响度范围内的演奏音频，每种乐器录制了涵盖其典型演奏范围的样本。通过 Librosa 工具提取多维声学特征。

对每个特征计算其统计量，包括均值、标准差、最大值、最小值、偏度和峰度，这些统计量提供了对音频特征分布的更深刻理解。最终提取的特征是上述每一类特征的统计特征拼接，每个音频文件都会被转化为一个 217 维特征向量。

3.2 降维聚类算法的选择与优化

为可视化高维特征分布、提升聚类效果与效率，对比了 PCA、t-SNE 与 UMAP 三种降维方法。实验表明，UMAP 在保持局部与全局结构平衡性上表现最优。然而，初步降维结果显示三类乐器的样本仍存在部分重叠区域，如图 3(a)和(c)所示。

针对特征冗余问题，提出一种分块赋权策略，将特征划分为四段，采用贝叶斯优化框架动态调整各段权重。目标函数定义为聚类结果与真实标签的调整兰德指数 ARI。优化过程中，权重搜索空间设定为[0.5,5.5]，经 500 次迭代后，获得最优权重组合

$$\omega^* = [1.2257, 0.6168, 1.4417, 0.6469]$$

分别对应 MFCC、色度特征、光谱特征及谐波与噪声比的权重系数。赋权后 UMAP 投影显示三类乐器聚类区域分离度显著提升，ARI 值从

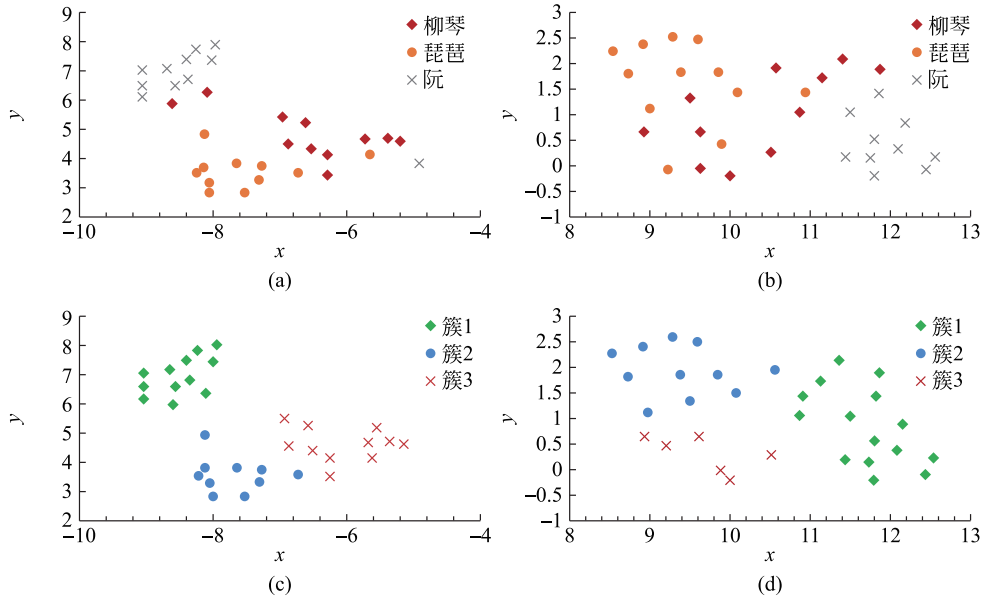

图 3 三类乐器声赋权前后的聚类效果图

(a) 原分类标签；(b) 原分类标签；(c) 聚类效果；(d) 聚类效果

0.3610 提高至 0.6661，如图 3(b)和(d)所示。

通过对比未赋权与赋权后的聚类效果，我们进一步验证了特征优化策略在音色区分中的有效性。如图 4 所示，图 4(a)和(c)为特征赋权后的降维模型的降维结果及其聚类结果图。可见柳琴与阮两种弦乐器的声学中心较为接近，簇间间距针较不明显，但簇间分界线清晰。合成弦乐器音与自然音声学中心较远，簇间距较大。对该降维结果进行 k-means 聚类，聚类结果与原分类标签完全匹配，ARI=1.0000。而对于未赋权优化的降维算法，如图 4(b)和(d)所示，合成音与自然音降维后的簇间距相对较小，界限不完全闭合。聚类后，ARI=0.9050，不能完全区分合成音与自然音。

以上结果说明，对合成音与自然音的混合数据集，未赋权模型的聚类结果显示两类样本存在一定重叠，表明传统方法难以完全区分模式相似

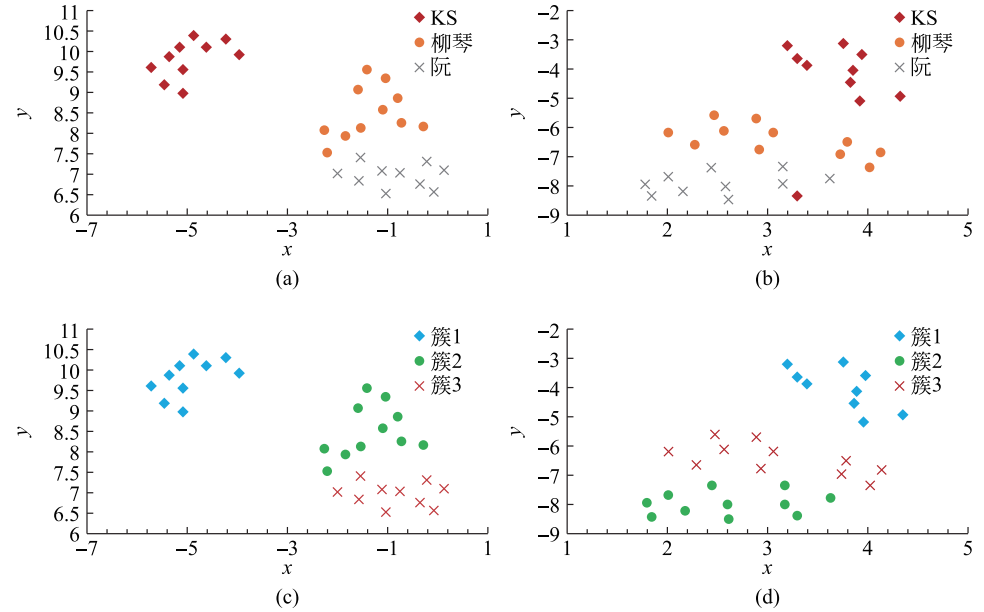

图 4 自然声与合成声赋权前后的聚类效果图

(a) 原分类标签；(b) 原分类标签；(c) 聚类效果；(d) 聚类效果

的自然音与合成音的信号。而采用预训练权重的赋权优化模型，呈现完全分离状态，证明算法对合成音与自然音的细微差异具有更强的鉴别能力。

3.3 残留重叠样本分析

尽管赋权优化提升了区分度，仍存在少量样本交叉现象。其主要成因在于声学参数的固有重叠：实验严格控制的频率与响度一致性设计，导致不同乐器的稳态声学特征呈现高度相似性。例如，柳琴与琵琶在相同基频下，其前三阶谐波能量差异小于 3dB，使得赋权策略难以完全解耦此类共性特征。此外，三类乐器均基于弦-共鸣腔耦合振动原理，其瞬态响应与稳态频谱存在物理规律性重叠。实验数据显示，柳琴与阮的衰减时间差异仅为 15ms，低于算法分辨率阈值。

3.4 声学特征区分度分析

赋权优化结果表明，权重 1.4417 的光谱特征与权重 1.2257 的 MFCC 是音色区分的核心维度，其主导性源于两者对频域能量分布及稳态频谱特性的精准刻画。光谱特征通过量化高频泛音占比、谐波集中度及频带能量差异，有效放大了柳琴、琵琶与阮的物理振动特性差异，如柳琴频谱质心较阮高 180Hz。MFCC 则通过 Mel 滤波器组压缩高频噪声，强化了共振峰分布的类间区分性。而色度特征与谐波比因音高重叠及实验噪声控制，贡献度受限。

3.5 合成优化

人工合成音通过预设参数模拟自然音特性，但其声学特征仍存在谐波结构偏差、瞬态响应简化的固有缺陷。强化学习通过最大化与自然音的声学特征相似度，动态调整合成参数，生成高度逼近自然音的合成信号。

在本部分，选用经过赋权优化后的降维聚类模型作为测试模型，该模型在上一部分已被证明具有较强的区分自然弦乐器声与 Karplus-Strong 算法合成的合成弦乐器声能力。自然声样本用一组 10 段阮拨弦声，在上一组实验中该样本组的声学中心距合成音相对较远（图 4）。将其与单个优化后合成音进行聚类测试。如图 5 所示，图(a)和(c)为人耳测试调参合成的合成音降维与聚类结果，图(b)和(d)为强化学习经 500 次迭代学习后的合成音结果。

实验结果表明，在分析人为设定参数的合成音时，赋权优化后的聚类模型声学中心间距 3169.2，ARI=1.0000，能够完全区分人工合成的近似自然音，如图 5(c)所示；而强化学习调参生成的合成音则导致聚类失效，声学中心间距 915.3，ARI=0.2877，如图 5(d)所示。

此结果再次展示了赋权模型对人工合成音的较强鉴别能力；同时，强化学习调参合成的音频信号能够有效规避基于静态特征的聚类算法，这一现象体现了该算法在声波仿真合成中的优势。

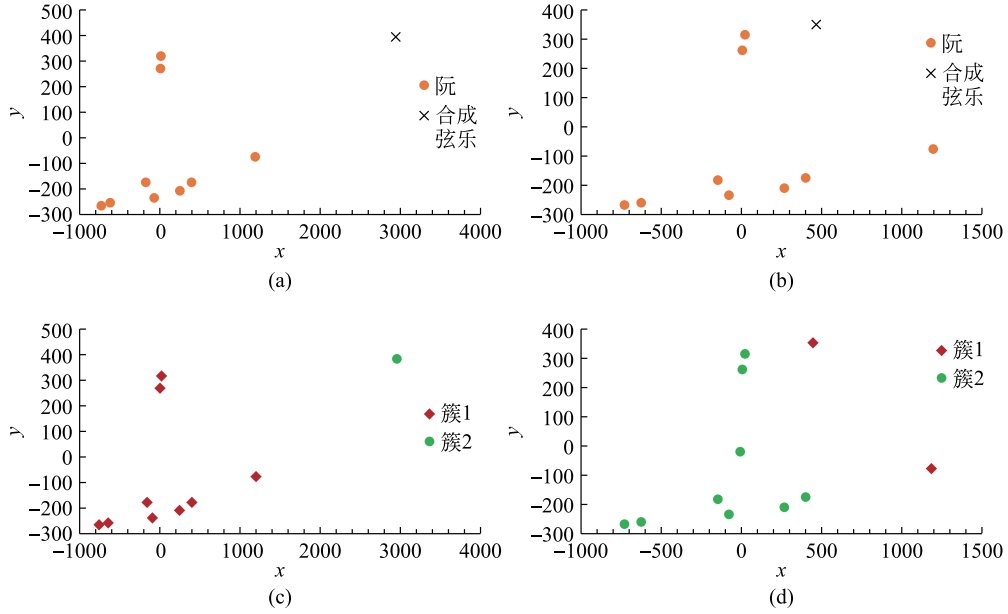

图 5 优化前后合成音的聚类效果图

(a) 原分类标签；(b) 原分类标签；(c) 聚类效果；(d) 聚类效果

4 结论

(1) 提出基于降维-聚类的合成音鉴别方法分块化贝叶斯赋权策略,通过概率模型动态筛选高维声学特征中的关键判别维度,提升合成音与自然音鉴别能力。

(2) 优化了合成音鉴别的计算效率,通过运用降维聚类结合赋权策略,在保证精度的同时降低计算复杂度。

(3) 通过强化学习最大化与自然音的声学特征相似度,动态调整合成参数,得到能够生成高度逼近自然音的合成信号的优化算法。

(4) 未来研究可进一步探索如何通过引入时间序列建模或物理声学方程来增强动态特征的表征能力,特别是针对瞬态响应与动态衰减特性等尚未充分建模的方面。

致谢:本项工作受无锡市科学技术局、无锡市滨湖区人民政府、无锡市滨湖区科学技术局提供的"未来技术太湖创新基金"项目资金支持,在此一并予以感谢。

参 考 文 献

[1] SMITH J O. Physical audio signal processing[M]. W3K Publishing, 2010.

[2] MCADAMS S, WINSBERG S. Caractérisation du timbre des sons complexes. Ⅱ. Analyses acoustiques et quantification psychophysique[J]. Journal de Physique IV, 1999, 9(PR8): 1-12.

[3] KARPLUS K, STRONG A. Digital synthesis of plucked string and drum timbres[J]. Computer Music Journal, 1983, 7(2): 43-55.

[4] ZÖLZER U. DAFX: Digital Audio Effects[M]. Hoboken: Wiley, 2011.

[5] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357-366.

[6] PEETERS G. A large set of audio features for sound description[J]. IRCAM Technical Report, 2004.

[7] MCINNES L, HEALY J, MELVILLE J. UMAP: Uniform manifold approximation and projection for dimension reduction[J]. arXiv: 1802.03426, 2018.

[8] HARTIGAN J A, WONG M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1): 100-108.

[9] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.

[10] SNOEK J, LAROCHELLE H, ADAMS R P. Practical Bayesian optimization of machine learning algorithms[J]. Advances in Neural Information Processing Systems, 2012, 25: 2951-2959.

[11] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8(3-4): 229-256.

[12] ENGEL J, RESNICK C, ROBERTS A, et al. DDSP: Differentiable digital signal processing[J]. International Conference on Learning Representations.

■