

群体皮电信号辅助的音乐情绪流识别

甄祥麟¹, 计晓斐^{1*}, 刘银华^{2*}

¹青岛大学计算机科学技术学院, 山东 青岛

²青岛大学自动化学院, 山东 青岛

收稿日期: 2026年9月19日; 录用日期: 2026年10月16日; 发布日期: 2026年10月28日

摘要

动态音乐情绪识别旨在刻画音乐推进过程中连续变化的效价-唤醒度轨迹。现有方法主要依赖音频和群体平均主观标签, 对音乐诱发的听者生理响应利用不足。皮电活动(Electrodermal Activity, EDA)反映汗腺活动引起的皮肤电学性质变化, 可表征听者在音乐聆听过程中与交感神经活动相关的连续生理反应, 为音乐情绪识别提供音频之外的生理证据。然而, 将多名听者的EDA序列直接汇总, 容易混合可重复的共同响应、听者差异与测量扰动。为此, 本文提出群体皮电辅助音乐情绪流网络(Collective Electrodermal Activity-Assisted Music Emotion Flow Network, CEA-Flow)。该网络将同一音乐刺激下的多听者EDA建模为无序时序集合, 通过不相交听者分组学习群体EDA共同响应, 并以听者差异信息估计响应的可靠性, 采用音频-EDA快、慢双时间尺度编码预测连续情绪轨迹。实验显示, 采用CEA-Flow模型预测的效价和唤醒度一致性相关系数分别为 0.201 ± 0.006 和 0.326 ± 0.005 , 平均一致性相关系数较仅音频特征模型预测提高0.048, 且优于歌曲错配和时间错位EDA对照。结果表明, 群体EDA共同响应能够为音频表征补充多听者在音乐刺激下的连续生理响应线索, 听者差异信息则有助于估计并调节这些线索在不同时间位置的可靠性。

关键词

动态音乐情绪识别, 皮电活动, 时间序列回归

Music Emotion Flow Recognition Assisted by Collective Electrodermal Activity

Xianglin Zhen¹, Xiaofei Ji^{1*}, Yinhua Liu^{2*}

¹College of Computer Science & Technology, Qingdao University, Qingdao Shandong

²School of Automation, Qingdao University, Qingdao Shandong

Received: September 19, 2026; accepted: October 16, 2026; published: October 28, 2026

*通讯作者。

文章引用: 甄祥麟, 计晓斐, 刘银华. 群体皮电信号辅助的音乐情绪流识别[J]. 软件工程与应用, 2026, 15(5): 749-759.
DOI: 10.12677/sea.2026.155070

Abstract

Dynamic music emotion recognition aims to characterize the continuously changing valence-arousal trajectory as music unfolds. Existing methods rely mainly on audio features and group-averaged subjective annotations, with limited use of the physiological responses induced by music. Electrodermal activity (EDA) reflects changes in the electrical properties of the skin caused by sweat gland activity. It can characterize continuous physiological responses associated with sympathetic activity during music listening and provide evidence beyond audio for music emotion recognition. However, directly aggregating EDA sequences from multiple listeners may mix repeatable common responses, listener differences, and measurement noise. To address this problem, we propose the Collective Electrodermal Activity-Assisted Music Emotion Flow Network (CEA-Flow). CEA-Flow models multi-listener EDA recorded under the same musical stimulus as an unordered set of temporal sequences. It learns the group EDA common response through disjoint listener groups, estimates response reliability from listener difference information, and predicts continuous emotion trajectories using Fast and Slow temporal encoding of audio and EDA. Experiments show that CEA-Flow achieves concordance correlation coefficients of 0.201 ± 0.006 for valence and 0.326 ± 0.005 for arousal. Its mean concordance correlation coefficient is 0.048 higher than that of the model using only audio features, and it outperforms the song-mismatched and time-shifted EDA controls. These results indicate that the group EDA common response can supplement audio representations with continuous physiological response cues shared by multiple listeners during music stimulation, while listener difference information helps estimate and regulate the reliability of these cues at different time positions.

Keywords

Dynamic Music Emotion Recognition, Electrodermal Activity, Time-Series Regression

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

音乐的节奏、音色、旋律、和声和曲式随演奏推进而变化,听者感知的情绪也呈现连续流动。效价-唤醒度(Valence-Arousal, V/A)空间分别描述情绪的愉悦程度和激活程度[1]。本文将这一连续轨迹称为音乐情绪流,并将动态音乐情绪识别(Dynamic Music Emotion Recognition, DMER)定义为从音乐及其相关信息中估计 V/A 轨迹的任务。与歌曲级分类相比,DMER 更关注变化方向、幅度和转折位置[2]。音乐情绪识别可广泛服务于情绪感知的音乐检索、推荐、片段定位和交互分析等场景[3]。

早期研究多使用能量、频谱、节奏、调性和音色等手工特征[4]。循环神经网络[5]、双向长短期记忆网络[6]进一步增强了双向时间依赖建模,多尺度注意力结构则用于聚合不同上下文范围[7]。自监督音乐模型 MERT 提供了丰富的逐帧音频表征[8],但现有 DMER 方法仍主要依赖音频,并常以单一时间结构处理不同变化速度的音频元素。

DMER 标签通常由多名听者的连续评分聚合得到,这虽然能够提供统一监督,却压缩了听者差异,也未直接呈现音乐刺激下的群体生理响应。多模态音乐情绪研究表明,不同信息源能够提供重要的互补情绪线索[9]。皮电活动(Electrodermal Activity, EDA)反映与交感神经活动相关的皮肤电学变化[10],其快

相(Phasic)和慢相(Tonic)成分可补充不同时间范围的生理信息。已有研究利用多尺度卷积提取 EDA 模式[11], 也有工作融合音乐与 EDA 进行大规模情绪识别[12]。然而, 对同一音乐下多听者的可重复 EDA 响应、听者差异及其可靠性仍缺少系统研究。

本文将同一音乐刺激下的多名听者 EDA 序列定义为群体皮电信号, 并研究如何分离其中可重复的群体 EDA 共同响应与听者差异信息, 使其分别提供连续生理线索和可靠性估计。为此, 本文提出群体皮电辅助音乐情绪流网络(Collective Electrodermal Activity-Assisted Music Emotion Flow Network, CEA-Flow)。模型通过相同歌曲不相交听者子集的一致性学习提取共同响应, 将听者差异转换为可靠性, 并以快速(Fast)/慢速(Slow)双时间尺度编码组织音频和 EDA 信息; 音乐构成主预测通路, EDA 通过有界残差提供补充。

本文的主要工作为: (1) 提出跨听者分组一致性学习, 提取群体 EDA 共同响应, 并用错配和错位序列检验其歌曲相关性与时间对应性; (2) 设计可靠性控制的音频-EDA 双时间尺度融合模型, 以群体 EDA 共同响应补充生理线索、以听者差异信息调节注入强度; (3) 通过表征验证、仅 EDA 分析、主模型比较、负对照和消融实验评估各组成部分。

2. 问题定义与研究方法

2.1. 任务定义

设一首歌曲在统一时间网格上的音乐表示为

$$\mathbf{X}^a = [\mathbf{x}_1^a, \dots, \mathbf{x}_T^a]^T \in \mathbb{R}^{T \times D_a}, \quad (1)$$

同一音乐刺激下共有 N 条有效 EDA 序列, 记为无序集合

$$\mathcal{E} = \left\{ \mathbf{E}^{(n)} \right\}_{n=1}^N, \quad \mathbf{E}^{(n)} \in \mathbb{R}^{T \times D_e}. \quad (2)$$

连续标签为 $\mathbf{Y} = [\mathbf{y}^v, \mathbf{y}^a] \in \mathbb{R}^{T \times 2}$, 两个通道依次表示效价和唤醒度。 $\mathbf{m}^y \in \{0, 1\}^T$ 表示监督有效帧, $\mathbf{M}^e \in \{0, 1\}^{N \times T}$ 表示不同听者的 EDA 有效位置。模型学习

$$\hat{\mathbf{Y}} = f_{\theta}(\mathbf{X}^a, \mathcal{E}, \mathbf{M}^e), \quad \hat{\mathbf{Y}} = [\hat{\mathbf{y}}^v, \hat{\mathbf{y}}^a] \in \mathbb{R}^{T \times 2}. \quad (3)$$

本文训练与推理均使用同步的音频和 EDA 输入。音频描述情绪刺激源, EDA 描述刺激后的群体生理响应, 二者在模型中承担不同角色。

2.2. CEA-Flow 总体结构

CEA-Flow 由音乐特征流、EDA 辅助流、有界残差融合和独立 V/A 动态输出头组成。音乐流从冻结 MERT 表征中学习 Fast 与 Slow 两类时间信息。EDA 辅助流先使用共享听者编码器处理每条记录, 再通过掩码集合注意力模块分离群体 EDA 共同响应与听者差异信息; 共同响应同样被分为 Fast 与 Slow 两个时间分支。最后, 由听者差异信息估计的可靠性限制 EDA 补充线索的注入幅度, 融合表示分别输入效价头和唤醒度头。整体模型结构如图 1 所示。

2.3. 双时间尺度音乐编码

音频经 24 kHz 重采样后输入冻结的 MERT-v1-330M。对齐到 2 Hz 后, 得到 24 层逐帧隐藏状态 $\mathbf{H}^{(l)} \in \mathbb{R}^{T \times 1024}$ 。本文使用可学习但归一化的层权重聚合这些表征:

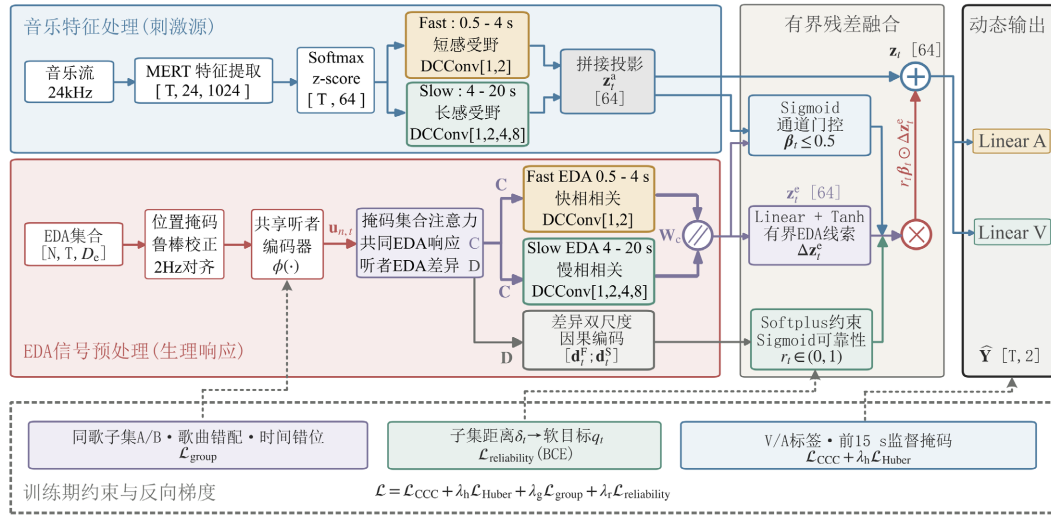


Figure 1. Overall framework of CEA-Flow

图 1. CEA-Flow 总体结构

$$\mathbf{x}_t^a = \mathbf{P}_a \sum_{l=1}^{24} \alpha_l \mathbf{h}_t^{(l)}, \quad \alpha_l = \frac{\exp(q_l)}{\sum_{j=1}^{24} \exp(q_j)}. \quad (4)$$

训练集逐通道统计量用于标准化 $\mathbf{x}_t^a$, MERT 参数保持冻结。Fast 音频编码器 $\mathcal{F}_a$ 由短感受野因果膨胀卷积构成, 关注约 0.5~4 s 内的局部能量、起音和节奏变化; Slow 音频编码器 $\mathcal{S}_a$ 使用长感受野因果膨胀卷积整合约 4~20 s 的上下文:

$$\begin{aligned} \mathbf{Z}^{a,F} &= \mathcal{F}_a(\mathbf{X}^a), & \mathbf{Z}^{a,S} &= \mathcal{S}_a(\mathbf{X}^a), \\ \mathbf{z}_t^a &= \mathbf{W}_a [\mathbf{z}_t^{a,F}; \mathbf{z}_t^{a,S}]. \end{aligned} \quad (5)$$

2.4. 跨听者分组一致性学习

对歌曲 i 的有效听者集合 $\mathcal{N}_i$, 每次训练随机采样两个互不重叠的子集 $\mathcal{N}_i^A$ 和 $\mathcal{N}_i^B$ 。两组 EDA 经过共享编码器和集合聚合器, 再对有效时间位置进行注意力池化, 得到歌曲级表示 $\mathbf{g}_i^A$ 与 $\mathbf{g}_i^B$ 。同一歌曲的两组表示保留可重复结构, 因此定义

$$\mathcal{L}_{\text{same}} = d(\mathbf{g}_i^A, \mathbf{g}_i^B), \quad (6)$$

其中 $d(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 + \varepsilon}$ 表示余弦距离。由于仅最小化同歌距离会使模型学习所有歌曲共有的时间模板, 为此, 从其他歌曲选取长度匹配表示 $\mathbf{g}_j^-$, 并对当前歌曲构造至少 8 s 的循环时间错位表示 $\tilde{\mathbf{g}}_i$ 。对比项为

$$\mathcal{L}_{\text{ctrl}} = [m + d(\mathbf{g}_i^A, \mathbf{g}_i^B) - d(\mathbf{g}_i^A, \mathbf{g}_j^-)]_+ + [m + d(\mathbf{g}_i^A, \mathbf{g}_i^B) - d(\mathbf{g}_i^A, \tilde{\mathbf{g}}_i)]_+. \quad (7)$$

其中, $m > 0$ 为间隔参数, $[x]_+ = \max(0, x)$ 。

同歌一致性和对照区分共同组成

$$\mathcal{L}_{\text{group}} = \mathcal{L}_{\text{same}} + \mathcal{L}_{\text{ctrl}}. \quad (8)$$

这一过程不使用 V/A 标签定义“共同性”，从而把跨听者可重复性与情绪任务相关性分开检验。

2.5. 跨听者 EDA 集合编码

EDA 记录先经过异常位置掩码、逐听者鲁棒中心校正和训练集尺度标准化。共享编码器 ϕ 将第 n 名听者在时刻 t 的输入映射为 $\mathbf{u}_{n,t}$ 。考虑到同一歌曲的记录顺序没有语义，集合聚合采用置换不变形式[13]：

$$\mathbf{c}_t = \sum_{n=1}^N w_{n,t} \mathbf{u}_{n,t}, \quad w_{n,t} = \frac{M_{n,t}^e \exp(s_{n,t})}{\sum_{j=1}^N M_{j,t}^e \exp(s_{j,t}) + \varepsilon}. \quad (9)$$

其中， $s_{n,t}$ 是共享评分器输出的标量注意力分数， $\varepsilon > 0$ 用于数值稳定。 $\mathbf{c}_t$ 表示群体 EDA 共同响应。听者差异信息采用加权绝对偏差表示：

$$\mathbf{d}_t = \sum_{n=1}^N w_{n,t} |\mathbf{u}_{n,t} - \mathbf{c}_t|. \quad (10)$$

绝对值按向量元素计算。将逐时刻表示堆叠为 $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_T]^T$ 和 $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_T]^T$ ，分别表示整首歌曲的群体 EDA 共同响应和听者差异信息。

图 2(a) 展示按预设中位代表性规则从训练集选取的歌曲及其发布版 EDA 记录；图 2(b) 展示鲁棒标准化后的听者轨迹，以及两个不相交听者子集的中心轨迹；图 2(c) 采用跨听者中位轨迹和基于四分位距(IQR)构造的离散带，直观呈现群体响应的集中趋势与听者间离散程度。这些描述性统计量仅用于数据可视化；模型实际使用式(9)和式(10)定义的注意力加权共同响应与加权绝对偏差。浅色背景表示不参与监督损失计算的前 15 s。

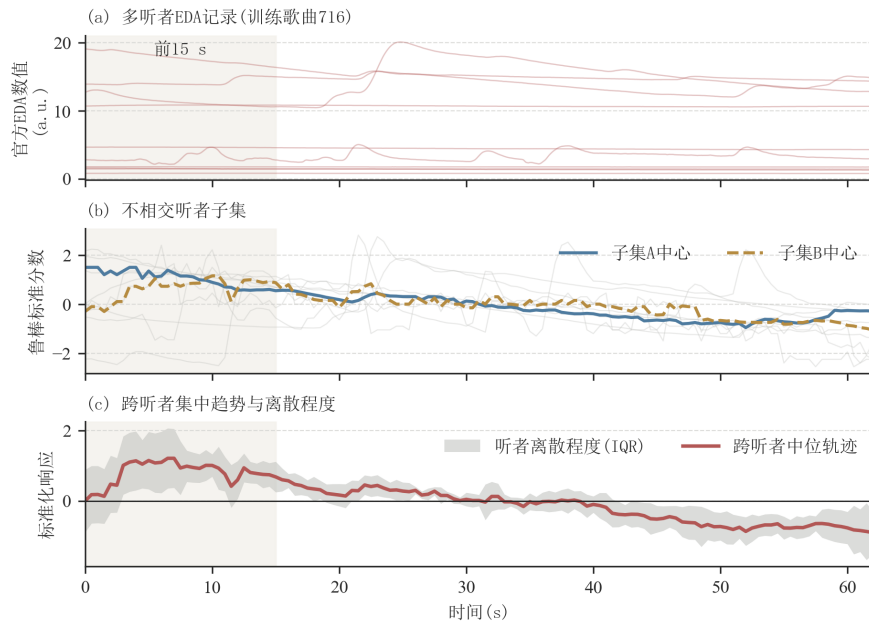


Figure 2. Multi-listener EDA and cross-listener descriptive statistics

图 2. 多听者 EDA 及其跨听者描述性统计

2.6. EDA Fast/Slow 表征

群体 EDA 共同响应分别进入 Fast 与 Slow 路径。Fast 路径使用短感受野因果膨胀卷积，关注快速上

升、峰值和恢复；Slow 路径使用较长膨胀率与因果平滑，描述较稳定的基线变化：

$$\begin{aligned} \mathbf{Z}^{e,F} &= \mathcal{F}_e(\mathbf{C}), \quad \mathbf{Z}^{e,S} = \mathcal{S}_e(\mathbf{C}), \\ \mathbf{z}_t^e &= \mathbf{W}_c \left[\mathbf{z}_t^{e,F}; \mathbf{z}_t^{e,S} \right]. \end{aligned} \quad (11)$$

听者差异序列采用对应的两个因果时间编码器：

$$\mathbf{D}^F = \mathcal{F}_d(\mathbf{D}), \quad \mathbf{D}^S = \mathcal{S}_d(\mathbf{D}). \quad (12)$$

其中， $\mathbf{z}_t^{e,F}$ 和 $\mathbf{z}_t^{e,S}$ 表示两个时间尺度上的群体 EDA 共同响应， $\mathbf{d}_t^F$ 和 $\mathbf{d}_t^S$ 是 $\mathbf{D}^F$ 和 $\mathbf{D}^S$ 在时刻 t 的行向量。EDA 与音频分支采用相近的功能性时间范围，以便在统一时间网络上交换局部和慢变信息。

2.7. 群体 EDA 共同响应与可靠性控制融合

听者差异信息通过单调函数转换为(0, 1)范围的逐时刻标量可靠性。首先使用 Softplus 将自由参数 θ_t 转换为非负权重：

$$\mathbf{a}_t = \text{softplus}(\theta_t), \quad r_t = \sigma\left(b_r - \mathbf{a}_t^T \left[\mathbf{d}_t^F; \mathbf{d}_t^S \right]\right). \quad (13)$$

其中， $\sigma(\cdot)$ 为 Sigmoid 函数， $\text{softplus}(x) = \log(1 + \exp(x))$ 并按元素作用于向量。由于 $\mathbf{a}_t \geq \mathbf{0}$ ，在其他分量不变时，任一听者差异分量增大都不会提高 r_t 。群体 EDA 共同响应生成有界补充线索 $\Delta \mathbf{z}_t^e = \tanh(\mathbf{W}_e \mathbf{z}_t^e)$ 。为保持音乐主导，使用上界为 $\beta_{\max} = 0.5$ 的通道级门控：

$$\beta_t = \beta_{\max} \sigma\left(\mathbf{W}_g \left[\mathbf{z}_t^a; \mathbf{z}_t^e \right] + \mathbf{b}_g\right), \quad (14)$$

$$\mathbf{z}_t = \mathbf{z}_t^a + r_t \beta_t \odot \Delta \mathbf{z}_t^e. \quad (15)$$

最终由两个独立线性头输出

$$\left(\hat{\mathbf{y}}^V\right)_t = \mathbf{w}_V^T \mathbf{z}_t + b_V, \quad \left(\hat{\mathbf{y}}^A\right)_t = \mathbf{w}_A^T \mathbf{z}_t + b_A. \quad (16)$$

该设计使群体 EDA 共同响应提供连续生理响应线索，听者差异信息控制其可信程度，避免低一致性的 EDA 信息过度影响情绪预测。

2.8. 优化目标

主任务在每首歌曲的有效监督帧上计算一致性相关系数(Concordance Correlation Coefficient, CCC) [14]。设 ρ_c 表示 CCC，则

$$\mathcal{L}_{\text{CCC}} = 1 - \frac{1}{2} \left[\rho_c(\hat{\mathbf{y}}^V, \mathbf{y}^V) + \rho_c(\hat{\mathbf{y}}^A, \mathbf{y}^A) \right]. \quad (17)$$

使用 Huber 损失约束逐帧误差，并通过分组一致性与对照区分学习 EDA 表征。可靠性目标由两个听者子集在时刻 t 的距离 δ_t 构造为 $q_t = \exp(-\kappa \delta_t)$ ，其中 q_t 在可靠性分支反向传播时停止梯度。采用二元交叉熵使 r_t 逼近 q_t ，使子集越一致的位置获得越高可靠性。总损失为

$$\mathcal{L} = \mathcal{L}_{\text{CCC}} + \lambda_h \mathcal{L}_{\text{Huber}} + \lambda_g \mathcal{L}_{\text{group}} + \lambda_r \mathcal{L}_{\text{reliability}}. \quad (18)$$

取 $\lambda_h = 0.2$ 、 $\lambda_g = 0.1$ 、 $\lambda_r = 0.05$ ，权重依据训练集和验证集确定。

3. 实验设置

3.1. 数据集与组约束划分

PMEmo 数据集包含 794 个流行音乐副歌片段、动态 V/A 标注和 7960 条同步 EDA 记录[15]，其中

767 首具有动态标注。按时长筛选后保留 672 首，并采用 568/64/40 的组约束训练、验证和测试划分。序列统一到 2 Hz，前 15 s 标注不参与训练和评价，详细协议见表 1。

Table 1. PMEmo data and group-constrained experimental protocol
表 1. PMEmo 数据与组约束实验协议

项目	设置
原始歌曲/动态标注歌曲	794/767
多听者 EDA 记录	7960 条；每首 10 或 11 条
可用歌曲与划分	672；训练/验证/测试为 568/64/40
EDA/动态标签采样率	50 Hz/2 Hz
监督起点	15 s 之后
评价单位	先逐歌计算，再跨歌曲和随机种子汇总
随机种子	1、2、3、4、42

3.2. 数据预处理

音频离线提取冻结 MERT 特征，形成 $[T, 24, 1024]$ 张量并对齐到 2 Hz；标准化统计量仅由训练歌曲估计。EDA 按听者进行鲁棒中心校正和训练集尺度标准化，中心与尺度作为听者描述保留。变长序列采用右侧填充，并分别维护序列、监督和听者有效掩码。Fast/Slow 分解由第 2.6 节的因果编码器完成。

3.3. 跨听者共性与任务相关性验证

在验证集上为每首歌曲和每个随机种子构造两个不相交听者子集，比较同歌、长度匹配错配歌曲和至少 8 s 时间错位表示的余弦相似度，并统计配对差值、95%置信区间、跨子集检索率和有效秩，以检验跨听者可重复性及表征退化。

随后冻结 EDA 编码器，仅在训练集拟合岭回归探测器，并在验证集逐歌预测 V/A；标签置换和时间错位结果用于检验任务相关性。

3.4. 对比方法

对比方法包括：(1) 仅音频模型(Audio-only)；(2) 固定统计、学习式、仅共同响应和仅差异信息四种 EDA-only 模型；(3) Audio + EDA-common、CEA-Flow 及歌曲错配、时间错位两个多模态对照。所有方法共享数据划分、监督掩码和评价指标。

3.5. 超参数与环境配置

音乐与 EDA 分支的隐藏维度均为 64，丢弃率(Dropout)为 0.2。模型使用 AdamW 优化器，初始学习率为 3×10^{-4} ，权重衰减为 10^{-3} ，批量大小为 8，最多训练 100 轮。验证集平均 CCC 用于学习率调度和早停，学习率调度与早停的耐心值分别为 4 和 12，梯度范数裁剪阈值为 1.0。全部实验在 NVIDIA RTX 4070 12 GB GPU 上使用混合精度运行，并采用固定随机种子 1、2、3、4 和 42。

3.6. 评价指标与统计方法

采用 CCC、皮尔逊相关系数(Pearson Correlation Coefficient, PCC)和均方根误差(Root Mean Square Error, RMSE)。CCC 同时评价相关性以及均值和尺度的一致性，PCC 关注轨迹形状，RMSE 衡量绝对误差。

所有指标先对每首歌曲分别计算，再对歌曲取平均；表中报告五个种子的均值与标准差。CEA-Flow 与基线采用相同随机种子配对，报告配对差值、基于 t 分布的 95%置信区间。

4. 实验结果与讨论

4.1. 跨听者分组一致性与任务相关性

表 2 和图 3 给出跨听者表征的前置验证。同歌不相交子集在 Fast/Slow 表征上均高于歌曲错配和时间错位对照，其中相对错位的相似度增量分别为 0.314 和 0.335，说明共同表示同时保留歌曲匹配与时间对齐结构。

跨子集 Top-1/Top-5 检索率达到 $0.219 \pm 0.018/0.538 \pm 0.019$ ，均高于机会水平，且有效秩为 18.1 ± 0.8 ，未出现表征塌缩。冻结岭回归探测器的 V/A CCC 为 $0.076 \pm 0.009/0.162 \pm 0.014$ ，明显高于标签置换和时间错位对照，表明该表征包含可预测的情绪相关信息，其中唤醒度关联更强。

Table 2. Validation of cross-listener EDA representations
表 2. 跨听者 EDA 表征验证

验证目标	指标	结果
可重复性	同歌相对错配的相似度增量	Fast 0.227 [0.222, 0.231]; Slow 0.223 [0.219, 0.226]
可重复性	同歌相对错位的相似度增量	Fast 0.314 [0.310, 0.318]; Slow 0.335 [0.331, 0.340]
表征辨识性	跨子集歌曲检索率	Top-1 0.219 ± 0.018 ; Top-5 0.538 ± 0.019
表征辨识性	有效秩	18.1 ± 0.8
任务相关性	冻结岭回归探测器	V-CCC 0.076 ± 0.009 ; A-CCC 0.162 ± 0.014
任务相关性	标签置换对照	V-CCC 0.007 ± 0.007 ; A-CCC -0.003 ± 0.017
任务相关性	时间错位对照	V-CCC 0.012 ± 0.003 ; A-CCC 0.022 ± 0.007

注：相似度增量表示同歌相似度减去相应对照相似度，方括号内为 95%置信区间。

图 3 进一步展示相似度增量、跨子集检索、有效秩及冻结探测器结果；单种子结果和 95%置信区间反映了上述趋势的稳定性。

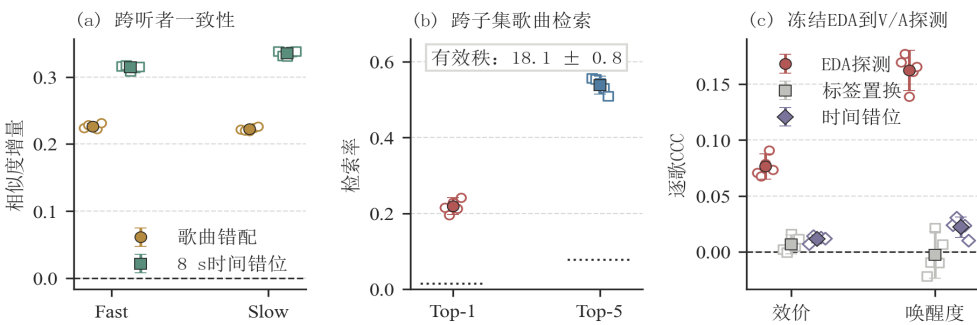


Figure 3. Cross-listener reproducibility and task relevance of EDA representations
图 3. 跨听者 EDA 表征的可重复性与任务相关性

4.2. EDA 表征能力

表 3 显示，学习式 EDA 集合编码的 V/A CCC 达到 $0.079 \pm 0.005/0.168 \pm 0.005$ ，优于固定统计。仅共

同响应模型接近完整 EDA 表征，而仅差异信息模型的平均 CCC 为 0.043 ± 0.009 ，说明共同响应更适合承载情绪相关线索，听者差异更适合作为可靠性依据。

仅 EDA 模型仍低于仅音频模型，符合音乐作为刺激源、EDA 作为受个体状态、设备和生理延迟影响的响应信号这一模态定位。

Table 3. Prediction performance of EDA-only representations
表 3. 仅 EDA 表征的预测性能

EDA 表征	V-CCC	A-CCC	平均 CCC
EDA-fixed	0.055 ± 0.005	0.125 ± 0.006	0.090 ± 0.005
EDA-learned	0.079 ± 0.005	0.168 ± 0.005	0.123 ± 0.005
EDA-common-only	0.073 ± 0.007	0.157 ± 0.005	0.115 ± 0.005
EDA-difference-only	0.026 ± 0.009	0.061 ± 0.010	0.043 ± 0.009

4.3. 与主要方法的比较

表 4 显示，CEA-Flow 的 V/A CCC 分别为 0.201 ± 0.006 和 0.326 ± 0.005 ，较仅音频模型的平均 CCC 提高 0.048，95%置信区间为[0.031, 0.064]，五个种子的差值均为正。PCC 和 RMSE 也同步改善，表明群体 EDA 信息同时提升了轨迹一致性和绝对误差表现。

相较 Audio + EDA-common，CEA-Flow 的平均 CCC 进一步提高 0.014，说明听者差异驱动的可靠性控制具有补充作用。歌曲错配和时间错位模型的平均 CCC 分别为 0.222 ± 0.007 和 0.227 ± 0.014 ，均接近仅音频模型且低于 CEA-Flow，表明增益依赖歌曲匹配和时间对齐的生理响应。唤醒度增益较大，而效价改善更多依赖慢变音频与 EDA 线索的共同作用。

Table 4. Main methods comparison
表 4. 主要方法比较

方法	V-CCC	A-CCC	V-PCC	A-PCC	V-RMSE	A-RMSE
Audio-only	0.158 ± 0.008	0.274 ± 0.010	0.302 ± 0.010	0.482 ± 0.010	0.104 ± 0.003	0.097 ± 0.002
Audio + EDA-common	0.188 ± 0.008	0.310 ± 0.006	0.352 ± 0.006	0.538 ± 0.006	0.098 ± 0.002	0.091 ± 0.002
CEA-Flow	0.201 ± 0.006	0.326 ± 0.005	0.379 ± 0.008	0.562 ± 0.005	0.095 ± 0.002	0.087 ± 0.002
Song-mismatched	0.164 ± 0.008	0.279 ± 0.006	0.311 ± 0.009	0.491 ± 0.006	0.103 ± 0.002	0.096 ± 0.002
Time-shifted	0.169 ± 0.013	0.285 ± 0.014	0.319 ± 0.014	0.499 ± 0.012	0.102 ± 0.004	0.095 ± 0.003

4.4. 消融实验

表 5 将消融分为信息角色与时间尺度两组。移除群体 EDA 共同响应后的 V/A CCC 降至 $0.174 \pm 0.009/0.289 \pm 0.009$ ，降幅大于移除可靠性控制，说明共同响应是主要补充信息，听者差异则用于抑制低一致性 EDA 的干扰。

时间尺度消融呈现不同侧重：移除 EDA Fast 主要降低 A-CCC，移除 EDA Slow 主要降低 V-CCC；音频 Slow 被移除时两个维度均明显下降，表明 EDA 能够补充但不能替代音乐上下文信息。

4.5. 音乐情绪流可视化

图 4 展示三种听者差异水平的代表样本。差异较小时，较高可靠性允许群体 EDA 共同响应对预测进

行有界修正，使模型更接近唤醒度局部峰值；差异较大时，可靠性降低，预测主要由音频分支决定。对于变化较缓的效价，Slow 音频与 EDA 线索共同调整轨迹基线。

Table 5. Ablation of information roles and temporal scales
表 5. 信息角色与时间尺度消融

方法	V-CCC	A-CCC	V-PCC	A-PCC	V-RMSE	A-RMSE
CEA-Flow	0.201 ± 0.006	0.326 ± 0.005	0.379 ± 0.008	0.562 ± 0.005	0.095 ± 0.002	0.087 ± 0.002
w/o group EDA common response	0.174 ± 0.009	0.289 ± 0.009	0.328 ± 0.009	0.505 ± 0.007	0.101 ± 0.002	0.094 ± 0.002
w/o reliability	0.188 ± 0.006	0.310 ± 0.009	0.352 ± 0.009	0.538 ± 0.008	0.098 ± 0.002	0.091 ± 0.002
w/o EDA Fast	0.198 ± 0.007	0.305 ± 0.009	0.371 ± 0.010	0.533 ± 0.009	0.096 ± 0.003	0.091 ± 0.003
w/o EDA Slow	0.181 ± 0.006	0.320 ± 0.006	0.341 ± 0.006	0.553 ± 0.006	0.100 ± 0.001	0.088 ± 0.001
w/o Audio Fast	0.194 ± 0.011	0.297 ± 0.011	0.365 ± 0.010	0.521 ± 0.010	0.097 ± 0.003	0.092 ± 0.003
w/o Audio Slow	0.166 ± 0.010	0.286 ± 0.011	0.316 ± 0.011	0.503 ± 0.009	0.102 ± 0.003	0.095 ± 0.003

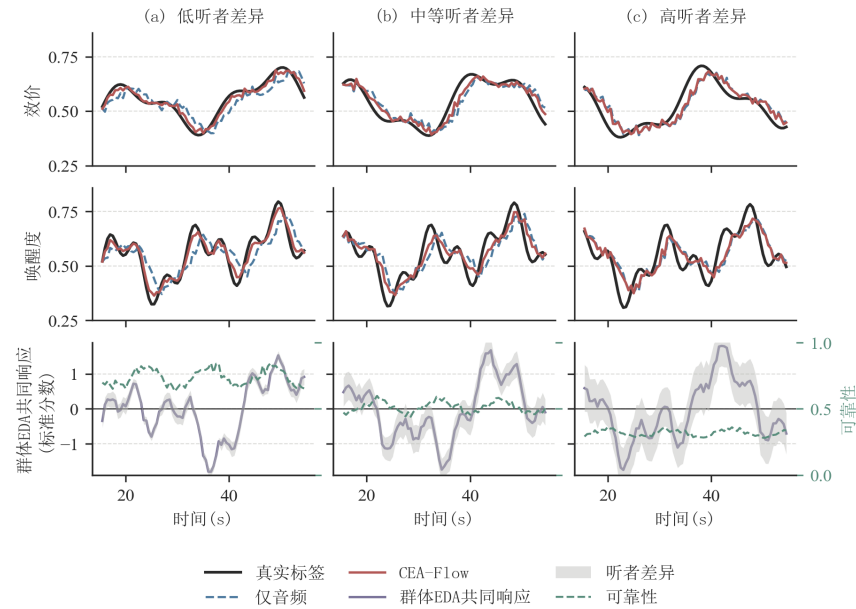


Figure 4. Representative music emotion-flow predictions at three listener-difference levels
图 4. 三种听者差异水平下的代表性音乐情绪流预测

5. 结论

本文提出 CEA-Flow 模型，从群体皮电信号中分离可重复的群体 EDA 共同响应与听者差异信息，并分别用于补充生理线索和估计可靠性。模型以 Fast/Slow 双时间尺度组织音频和 EDA 表征，其平均 CCC 较仅音频模型提高 0.048；负对照和消融进一步说明了正确刺激 - 响应对应、共同响应及可靠性控制的作用。结果表明，群体 EDA 可作为音频之外的补充信息，用于连续音乐情绪预测。后续研究将进一步考虑不同听者的生理响应延迟，增强模型对听者数量变化、部分 EDA 缺失和测量噪声的适应能力，并在更多具有连续情绪标注和同步生理信号的数据集上验证其跨数据集泛化能力。在此基础上，还可将脑电、心电图等生理信号纳入统一的可靠性控制框架，研究不同生理响应对效价与唤醒度动态的互补作用。

基金项目

青岛大学“系统科学+”联合研究计划项目(XT2024102) (省部级)。

参考文献

- [1] Russell, J.A. (1980) A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, **39**, 1161-1178. <https://doi.org/10.1037/h0077714>
- [2] Schubert, E. (2004) Modeling Perceived Emotion with Continuous Musical Features. *Music Perception*, **21**, 561-585. <https://doi.org/10.1525/mp.2004.21.4.561>
- [3] 康健, 王海龙, 苏贵斌, 等. 音乐情感识别研究综述[J]. 计算机工程与应用, 2022, 58(4): 64-72.
- [4] 陈晓鸥, 杨德顺. 音乐情感识别研究进展[J]. 复旦学报(自然科学版), 2017, 56(2): 136-148.
- [5] Weninger, F., Eyben, F. and Schuller, B. (2014) On-Line Continuous-Time Music Mood Regression with Deep Recurrent Neural Networks. 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 4-9 May 2014, 5412-5416. <https://doi.org/10.1109/icassp.2014.6854637>
- [6] Li, X., Xianyu, H., Tian, J., Chen, W., Meng, F., Xu, M., et al. (2016) A Deep Bidirectional Long Short-Term Memory Based Multi-Scale Approach for Music Dynamic Emotion Prediction. 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 20-25 March 2016, 544-548. <https://doi.org/10.1109/icassp.2016.7471734>
- [7] Ma, Y., Li, X., Xu, M., Jia, J. and Cai, L. (2017) Multi-Scale Context Based Attention for Dynamic Music Emotion Prediction. *Proceedings of the 25th ACM International Conference on Multimedia*, Mountain View, 23-27 October 2017, 1443-1450. <https://doi.org/10.1145/3123266.3123408>
- [8] Li, Y., Yuan, R., Zhang, G., et al. (2024) Mert: Acoustic Music Understanding Model with Large-Scale Self-Supervised Training. *International Conference on Learning Representations 2024*, Vienna, 7-11 May 2024, 12181-12204.
- [9] 韩东红, 孔彦茹, 展艺萌, 等. 音乐多模态数据情感识别方法的研究[J]. 东北大学学报(自然科学版), 2024, 45(6): 776-785, 792.
- [10] Boucsein, W. (2012) *Electrodermal Activity*. Springer Science & Business Media.
- [11] Ganapathy, N., Veeranki, Y.R., Kumar, H. and Swaminathan, R. (2021) Emotion Recognition Using Electrodermal Activity Signals and Multiscale Deep Convolutional Neural Network. *Journal of Medical Systems*, **45**, Article No. 49. <https://doi.org/10.1007/s10916-020-01676-6>
- [12] Yin, G., Sun, S., Yu, D., Li, D. and Zhang, K. (2022) A Multimodal Framework for Large-Scale Emotion Recognition by Fusing Music and Electrodermal Activity Signals. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **18**, 1-23. <https://doi.org/10.1145/3490686>
- [13] Zaheer, M., Kottur, S., Ravanbakhsh, S., et al. (2017) Deep Sets. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3394-3404.
- [14] Lin, L.I. (1989) A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, **45**, 255-268. <https://doi.org/10.2307/2532051>
- [15] Zhang, K., Zhang, H., Li, S., Yang, C. and Sun, L. (2018) The PMEmo Dataset for Music Emotion Recognition. *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, Yokohama, 11-14 June 2018, 135-142. <https://doi.org/10.1145/3206025.3206037>