基于自监督与知识蒸馏的轻量化声学场景分类

汪慧娟[1] 李宏滨[1]
WANG Huijuan LI Hongbin

摘 要 针对声学场景分类（ASC）任务中标注数据稀缺与边缘设备部署受限的双重挑战，文章提出了一种融合自监督学习（SSL）与知识蒸馏（KD）的创新框架。通过引入 HuBERT 自监督预训练模型，从大规模未标注的 TAU Urban Acoustic Scene 2024 数据中提取通用声学表征，并结合频谱混合风格（FMS）与自动设备脉冲响应（ADIR）增强策略，有效缓解设备域偏移问题。为实现边缘部署，设计双阶段轻量化架构：通过 HuBERT 教师模型生成软标签指导训练，采用 MobileNetV2 作为学生模型，在 DCASE 数据集上达到 81.3% 的准确率。结合量化感知训练（QAT）将模型压缩至 115.6 kB，同时，通过残差归一化层与深度可分离卷积的混合结构，在 TAU Urban 数据集上实现设备无关性分类，显著提升实际场景鲁棒性。

关键词 声学场景分类；深度学习；轻量化模型；自监督学习；知识蒸馏；量化感知训练

doi：10.3969/j.issn.1672-9528.2025.09.010

0 引言

声学场景分类（acoustic scene classification, ASC）技术演进聚焦于特征优化与模型泛化两大方向：早期依赖手工特征，如 MFCC、短时能量。2021 年，Prabakaran 等人 [1] 基于 MFCC 的语音处理研究，以及 2022 年高磊等人 [2] 结合短时能量与 MFCC 的联合特征分析，均是这一阶段的典型探索。随着深度学习发展，研究转向自动特征提取与鲁棒性优化，逐步推动 ASC 从人工规则驱动迈向数据驱动的泛化能力提升。

但受限于特征维度单一性，难以捕捉复杂场景的时频关联特性。为此，研究者引入时频联合分析手段，如 Gamma 谱图、深度散射光谱。2019 年，Zhang 等人 [3] 结合深度散射光谱（DSS）优化高频噪声鲁棒性。2020 年，Wang 等人 [4] 通过 Gamma 谱图增强中高频特征表达。然而，传统方法对设备噪声、声源混叠的适应性有限，学界转向端到端深度学习框架，2021 年，Alsayadi 等人 [5] 提出的基于 CTC、CNN-LSTM 与注意力机制的混合架构，通过联合优化特征提取与分类决策提升复杂场景下的抗干扰能力。

在模型架构层面，卷积神经网络（CNN）凭借局部特征提取优势成为声学场景分类的主流选择，研究者通过引入残差连接，2023 年，冯成立等人 [6] 提出的 DRCNN 模型；空洞卷积，2021 年，Tang 等人 [7] 基于上下文堆叠的空洞卷积架构等结构优化梯度传播与长时模式捕获，并采用多尺度特征融合策略增强复杂声学场景的建模能力。

值得注意的是，设备域偏移问题逐渐凸显：不同采集设备导致的声学特性差异显著降低模型泛化性能。现有解决方案多采用数据增强，如 Mixup、设备脉冲响应模拟，2020 年，Wei 等人 [8] 提出了一种结合 SpecAugment 和 Mixup 两种数据增强技术的音频分类方法。但依赖大量标注数据的特性限制了其在低资源场景的应用。

本研究针对声学场景分类中跨设备泛化性弱、数据标注成本高及边缘部署效率低的问题，提出融合自监督预训练与知识蒸馏的协同优化框架，通过 HuBERT 预训练提取通用声学表征，结合 MobileNetV2 量化蒸馏，实现 81.3% 分类准确率，模型压缩至 115.6 kB，参数量减少 25%。

1 自监督学习和知识蒸馏

1.1 自监督学习

本研究致力于在标注数据有限的条件下实现高精度声学场景分类，为此引入自监督语音表征学习方法。近年来，用于语音表示的自我监督学习框架，如 HuBERT 和 wav2vec2.0 已经证明了其在提取有用的语音音频特征和处理各种下游任务方面的强大能力。

HuBERT 相较于其他自监督音频模型，如 wav2vec 2.0、Data2vec，核心优势在于层次化表征与低资源适应性：通过迭代聚类生成稳定伪标签，首轮基于 MFCC，次轮基于中间特征，其低资源性能优于其他主流模型，有限标注数据下，词错误率较 wav2vec 2.0 降低 19%[9]，中文场景 CER 较

1. 太原师范学 山西晋中 030619

FBank 降低 30%[10]。

1.2 知识蒸馏

知识蒸馏（knowledge distillation）是一种模型压缩技术，通过训练小型“学生模型”模仿大型“教师模型”的输出分布，实现知识迁移。其核心在于利用教师模型的软标签而非硬标签，传递更丰富的类别间关系信息。如图 1 所示。

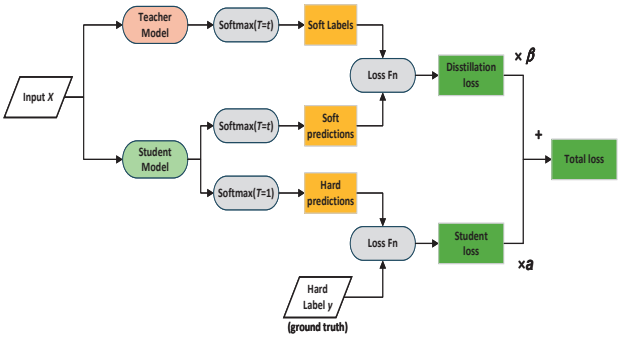

图 1 知识蒸馏流程图

教师模型经过带温度参数（T=t）的 Softmax 函数生成软标签，学生模型经过标准 Softmax（T=1）生成硬标签，经过与教师模型相同温度 T=t 的 Softmax 生成分类概率分布。

1.2.1 Softmax 函数

传统分类任务中，Softmax 将神经网络的原始输出（logits）映射为概率分布，标准 Softmax 函数为：

$$P_i=\frac{\exp(\boldsymbol{Z}_i)}{\sum_{j-1}^{C}\exp(\boldsymbol{Z}_j)} \tag{1}$$

式中：输入 logits 向量 $\boldsymbol{Z}=[z_1,z_2,...,z_C]\in\mathbb{R}^C$，其中 C 为类别数。

引入温度参数 T 后，Softmax 公式修正为：

$$P_i^T=\frac{\exp(\boldsymbol{Z}_i/T)}{\sum_{j-1}^{C}\exp(\boldsymbol{Z}_j/T)} \tag{2}$$

1.2.2 损失函数

知识蒸馏的损失计算通过双路径实现：教师模型输出的软标签与学生模型的软预测间计算 KL 散度损失，即传递类别间关联性，同时学生模型的硬预测与真实标签计算交叉熵损失、总损失为两部分的加权和。用公式表示为：

$$D_{\mathrm{KL}}(P\|Q)=\sum P_i\log\frac{P_i}{Q_i} \tag{3}$$

2 数据预处理与数据增强

2.1 数据集

TAU Urban Acoustic Scene 2022 移动开发数据集作为研究的训练集，包含由多城市移动设备采集的 230 350 条 /s 城市环境音频，涵盖机场、公交、地铁等 10 类声学场景硬标签。TAU Urban Acoustic Scene 2024 数据集基于同一数据集，提供 5%、10%、25%、50%、100% 比例的训练子集，旨在声学场景分类（ASC）系统中验证不同数据量下的性能表现。

2.2 数据增强

本研究针对声学场景分类中的数据稀疏性与设备域差异挑战，提出融合设备扰动和频谱增强的双阶段预处理框架，结合动态随机采样策略构建异构声学环境；在此基础上，通过蒸馏损失约束 MobileNetV2 学生模型与 HuBERT 教师模型的高层特征分布对齐，实现跨设备场景的强泛化建模。

2.2.1 自动设备脉冲响应 ADIR

ADIR 通过设备脉冲响应的参数化建模与动态卷积，将设备域偏移问题转化为可控的数据增强任务，其本质是在信号层面构造设备域的对抗样本，迫使模型忽略设备相关的卷积信道特性。从优化视角看，ADIR 可视为对设备域分布 $P(h)$ 的显式建模，其增强过程等价于在损失函数中引入设备不变性正则项：

$$L=_{\mathrm{ADIR}}=E_{h\sim p(h)}\ L_{\mathrm{CE}}\left(f(x\cdot h),h\right) \tag{4}$$

2.2.2 频谱混合风格 FMS

频谱混合风格（frequency-mix style, FMS）增强技术的核心在于通过随机频谱混合与风格扰动，破坏设备相关的局部频域特征分布，同时保留场景分类所需的全局声学语义。其数学实现可分为以下三个阶段：

（1）时域分解与幅度调制：打破设备相关的局部频带统计分布；

（2）相位扰动与频域插值：增加跨设备相位一致性约束；

（3）时域重构与风格正则化：抑制高频伪影引起的过拟合。

3 声学场景分类模型

针对声学场景分类任务中标注数据稀缺与边缘设备部署受限的双重挑战，本研究提出一种面向边缘设备声学场景分类的高效鲁棒模型，基于自监督学习与知识蒸馏协同优化框架设计。

3.1 教师模型 HuBERT

HuBERT（hidden-unit BERT）是一种基于 Transformer 架构的语音自监督预训练模型，其核心思想是通过聚类生成离散的语音单元作为伪标签，结合掩码语言建模任务进行训练。HuBERT 的网络架构由特征编码器、Transformer 编码器和聚类模块 3 部分组成，整体采用分阶段迭代训练策略 [11]。

3.1.1 特征编码器

特征编码器采用 7 层一维 -CNN 处理 16 kHz 原始语音，每层含 3 宽卷积核、512 通道，通过 padding=1 保持时序，步长逐层递减（10 到 2）实现多尺度特征压缩。结构通过不同时间分辨率的卷积操作，同时捕捉短时频谱细节和长时上下文依赖，为下游任务提供高维时序特征表达，用公式表

示为：

$$F^{l}=\mathrm{ReLU}\left(W^{l}*F^{l-1}+b^{l}\right) \tag{5}$$

式中：$l\in\{1,2,\cdots,7\}$ 表示层序号；$*$ 表示一维卷积操作；$W^{l}\in\mathbb{R}^{k\times C_{\mathrm{in}}\times 512}$ 表示可学习权重，其中 $C_{\mathrm{in}}=1$ 仅在首层输入时为单通道。

3.1.2 聚类模块

HuBERT 模型的聚类模块基于 MFCC 或模型中间特征生成连续向量，利用 K-means 聚类将其离散化为指定类别数的伪标签，通过模型训练与聚类中心更新的多轮迭代优化标签质量，如图 2 所示。

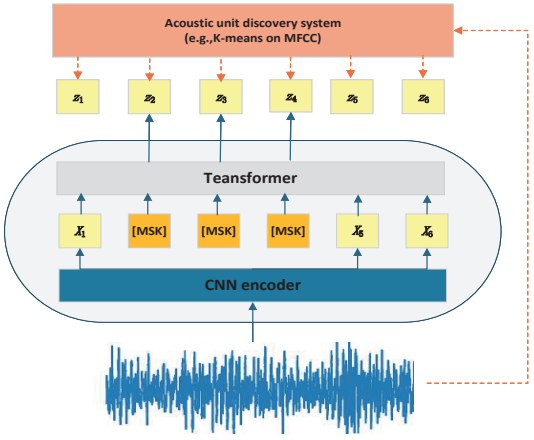

图 2 HuBERT 模型架构图

（1）聚类目标函数

假如输入语音帧序列为 $X=[x_1,x_2,\cdots,x_T]$，聚类模块的目标是将其映射为离散的隐藏单元序列 $Z=[z_1,z_2,\cdots,z_T]$，其中 $z_T\in\{1,2,\cdots,C\}$ 表示每个帧对应的类别编号。聚类过程通过 K-means 算法实现，其优化目标为：

$$\min_{\{\mu\}_{c=1}^{C}}\sum_{t=1}^{T}\left\|\boldsymbol{x}_t-\mu_{z_t}\right\|^2 \tag{6}$$

式中：μ_c 是第 c 个聚类中心；$\boldsymbol{x}_t$ 是输入特征。

（2）掩码预测的损失函数

$$L_m=-\sum_{t\in M}\log P\left(z_t\mid\mathrm{MaskedInput}\right) \tag{7}$$

式中：掩码位置集合 M 通过随机跨度掩码生成，预测概率通过余弦相似度计算。

$$P\left(z_t=c\right)=\frac{\exp\left(\mathrm{sim}\left(\boldsymbol{A}o_t,\boldsymbol{e}_c\right)/\tau\right)}{\sum_{c'=1}^{C}\exp\left(\mathrm{sim}\left(\boldsymbol{A}o_t,\boldsymbol{e}_{c'}\right)/\tau\right)} \tag{8}$$

式中：o_t 是 Transformer 编码器的输出；$\boldsymbol{A}$ 是投影矩阵，$\boldsymbol{e}_c$ 是第 c 类的嵌入向量；$\tau=0.1$ 为温度参数。

3.1.3 Transformer 编码器

Transformer 编码器由多层堆叠结构构成，每层含多头自注意力，捕捉全局依赖；以及和前馈网络，强化局部特征，结合残差连接与层归一化，有效缓解梯度消失、加速收敛并提升模型稳定性。

（1）残差连接与层归一化

每个子模块的输出通过以下步骤处理：

$$\mathrm{Output}=\mathrm{LayerNorm}\left(x+\mathrm{SubLayer}\left(x\right)\right) \tag{9}$$

多头自注意力子层：

$$x_{\mathrm{attn}}=\mathrm{LayerNorm}\left(x+\mathrm{MultiHeadAttention}\left(x\right)\right) \tag{10}$$

前馈神经网络子层：

$$x_{\mathrm{ffn}}=\mathrm{LayerNorm}\left(x_{\mathrm{attn}}+\mathrm{FFN}\left(x_{\mathrm{attn}}\right)\right) \tag{11}$$

式中：SubLayer 为子模块函数（自注意力或 FFN）；LayerNorm 使用可学习的缩放参数（γ）和平移参数（β）调整归一化后的分布。

（2）前馈神经网络

FFN 的典型实现（以 ReLU 激活为例）：

$$\mathrm{FFN}(x)=\max(0,x\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{12}$$

式中：$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为权重矩阵。

3.2 学生模型 MobileNetV2

MobileNetV2 作为轻量化卷积网络，采用倒置残差结构与深度可分离卷积，通过线性瓶颈层保留核心声学特征，在显著降低参数量的同时维持建模能力。在声学场景分类任务中，2020 年研究人员提出了一种名为 DS-FlexiNet 的改进模型。其核心基于 MobileNetV2 的深度可分离卷积模块，实现了效率和准确性的平衡 [12]，为低资源场景提供硬件友好的轻量化解决方案。

3.3 知识蒸馏过程

本文提出一种基于两阶段知识蒸馏的轻量化语音特征学习框架。

3.3.1 前端特征适配蒸馏

（1）输入音频经频谱混合风格（FMS）与自动设备脉冲响应（ADIR）增强后，生成多域混合特征。

$$x_{\mathrm{mix}}=\mathrm{FMS}\left(x_{\mathrm{clean}}\oplus\mathrm{noise}_{\mathrm{ADIR}}\left(\Delta f,\Delta\phi\right)\right) \tag{13}$$

式中：Δf 为幅度扰动；$\Delta\phi$ 为相位偏移。

（2）教师模型 HuBERT 的 CNN 编码器输出帧级特征 $\boldsymbol{Z}_T\in\mathbb{R}^{T\times 368}$，通过可学习投影矩阵 $\boldsymbol{W}_p\in\mathbb{R}^{768\times 384}$ 降维。

$$\boldsymbol{Z}_T'=\mathrm{LayerNorm}\left(\boldsymbol{W}_p\boldsymbol{Z}_T+\boldsymbol{b}_p\right) \tag{14}$$

（3）学生模型 MobileNetV2 前端网络同步提取 Fbank+CNN 特征，采用余弦相似度损失优化特征空间对齐。

$$L_{\mathrm{front}}=\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{\left\langle\boldsymbol{Z}_T'(i),\boldsymbol{Z}_s(i)\right\rangle}{\left\|\boldsymbol{Z}_T'(i)\right\|_2\cdot\left\|\boldsymbol{Z}_s(i)\right\|_2}\right) \tag{15}$$

该阶段冻结教师参数，通过 EMA 更新学生前端批归一化层，实现 2 倍时序下采样与特征维度压缩。

3.3.2 分层语义蒸馏

基于跨层注意力路由与自适应温度调节机制，实现深层次语言知识的迁移。

（1）跨层动态路由：计算学生第 j 层与教师各层 i 的注意力相似度，选取 Top-3 教师层作为监督源，构建多粒度知识传递路径。

（2）自适应温度蒸馏：动态调整 KL 散度的温度参数。

$$T(t)=T_{\max}\cdot e^{-\beta t}+T_{\min} \tag{16}$$

式中：β=0.01 控制衰减速率；初始 $T_{\max}$=8 为软化分布；后期 $T_{\min}$=2 为聚焦判别区域。

分层蒸馏损失为：

$$\mathcal{L}_{\mathrm{KD}}=\sum_{k=1}^{3}\frac{T_k^2}{N_k}\sum_{i=1}^{N_k}D_{\mathrm{KL}}\left(P_T^{(k)}\left(z_T^{(i)}/T_k\right)\middle\|P_S^{(k)}\left(z_S^{(i)}/T_k\right)\right) \tag{17}$$

（3）量化感知训练：在蒸馏后期引入权重量化模拟。

$$W_{\mathrm{quant}}=\mathrm{clip}\left(\left\lfloor\frac{W-\mu}{2\sigma}\ 255\right\rceil,0,255\right)\frac{2\sigma}{255}+\mu \tag{18}$$

联合优化目标：总损失函数融合特征对齐、知识蒸馏与任务监督。

$$L_{\mathrm{total}}=\alpha_1 L_{\mathrm{front}}+\alpha_2 L_{\mathrm{KD}}+\alpha_3 L_{\mathrm{CE}} \tag{19}$$

4 实验结果与分析

4.1 性能指标

分类准确率：基于设备无关测试集的宏观平均准确率（accuracy）。

设备鲁棒性：S4-S6 设备衰减率（%）衡量模型在未知设备（S4-S6）上的性能衰减程度；模型存储大小（kB）是衡量算法边缘端实用性的核心指标：较小的存储占用直接适配手机、IoT 等存储资源受限的嵌入式平台。

计算效率：每秒百万次乘加运算（MACs/s），模型参数量。

4.2 对比基准

本研究选取 3 类模型构建对比体系：DCASE 2024 基线 CP-Mobile 简化版，记为 BASE_I；EfficientNet-B0，复合缩放优化跨设备鲁棒性，记为 BASE_II；BEATs 预训练 MobileNetV2，自监督预训练与知识蒸馏对照，记为 BASE_III，与本模型 H_M 进行跨层级性能对比，验证设备适应性、量化效率及特征表达优势。

4.3 实验对比结果

本研究提出的 H_M 模型在参数量、计算效率与分类准确率 3 个维度超越基线系统：H_M 模型在参数量上较 BASE_I 减少 25%，MACs 运算量较 BASE_III 降低 13.3%，同时准确率提升至 81.3%，验证了其在高效率与高性能间的平衡优化，如表 1 所示。根据实验结果，H_M 模型在设备鲁棒性与存储效率上均展现显著优势：其 S4-S6 设备衰减率仅 8.2%，较最优基线 BASE_I 降低 41.8%，同时模型存储大小压缩至 115.6 kB 较 BASE_I 减少 9.5%，验证了模型架构对设备差异的强适应性及资源受限场景的部署可行性。如表 2 所示。

表 1 模型性能对比

模型	参数量 /10^6	MACs/10^7	准确率 /%
BASE_I	1.2	4.1	74.2
BASE_II	3.8	3.9	76.5
BASE_III	2.1	3.0	77.1
H_M	**0.9**	**2.6**	**81.3**

表 2 模型性能比较

模型	S4-S6 设备衰减率 /%	模型存储大小 /kB
BASE_I	14.1	127.8
BASE_II	19.8	384.6
BASE_III	15.6	194.2
H_M	**8.2**	**115.6**

5 结语

本研究创新性地提出一种两阶段动态知识蒸馏框架，实现 HuBERT 向 MobileNetV2 的高效知识迁移，在 TAU 2024 数据集声学场景分类任务中准确率达 81.3%，验证其对复杂声学环境的强鲁棒性；未来将融合多设备协同蒸馏与时频域解耦增强技术，优化低信噪比、多声源耦合场景的分类精度，推动轻量化声学模型在智能家居、工业监测等边缘场景的落地应用。

参考文献：

[1]PRABAKARAN D , SRIUPPILI S. Speech processing: MFCC based feature extraction techniques-an investigation[J]. Journal of physics conference series, 2021, 1717(1): 012009.

[2] 高磊 , 刘振奎 , 魏晓悦 , 等 . 铁路隧道二次衬砌敲击检查声音特征分析及智能识别 [J]. 铁道科学与工程学报 , 2022, 19(7): 1997-2004.

[3]ZHANG P Y, CHEN H T, BAI H C, et al. Deep scattering spectra with deep neural networks for acoustic scene classification tasks[J]. Chinese journal of electronics, 2019, 28(6): 1177-1183.

[4]WANG H L, ZOU Y X, CHONG D D. Acoustic scene classification with spectrogram processing strategies[EB/OL].(2020-07-06)[2025-06-26].https://doi.org/10.48550/arXiv.2007.03781.

[5]ALSAYADI H A, ABDELHAMID A A, HEGAZY I, et al. Arabic speech recognition using end-to-end deep learning[J]. IET signal processing, 2021, 15(8): 521-534.

[6] 冯成立 , 程雯 . 基于残差卷积神经网络的语音识别算法 [J].

基于骨骼动作识别的交叉注意多尺度时空 Transformer

高逸畅[1] 林哲煌[1]
GAO Yichang LIN Zhehuang

摘 要 近年来，Transformer 在计算机视觉各类任务中成效显著。但基于 Transformer 的方法在骨骼数据多尺度特征学习上存在局限，而多尺度时空特征蕴含着关键的全局与局部信息，这对于骨骼动作识别而言至关重要。为此，文章探索了骨骼序列在空间和时间维度上的多尺度特征表示，并提出了一种用于跨尺度特征融合的高效交叉注意力机制。此外，提出了一个多尺度特征提取和融合 Transformer（multi-scale feature extraction and fusion transformer, MFEF-Former），可以分为两种类型：（1）用于空间建模的 MFEF-SFormer，利用自注意力捕获关节间和身体部位间的相关性，然后利用交叉注意力进行多尺度空间特征融合，以建模关节和身体部位之间的相关性；（2）用于时间建模 MFEF-TFormer，利用自注意力捕获多尺度时间特征，并通过交叉注意力融合多尺度特征。这两个组件在一个双流网络中结合，并在两个大型数据集 NTU RGB+D 和 NTU RGB+D 12 上进行了评估。实验表明，文章所提方法在基于骨骼的动作识别方面优于其他基于 Transformer 的方法。

关键词 动作识别；Transformer；人体骨架；多尺度特征

doi：10.3969/j.issn.1672-9528.2025.09.011

0 引言

随着计算机视觉技术的快速发展，人类动作识别已成为一项日益流行和重要的任务，其应用涵盖交通运输、医疗、娱乐、教育、安全监控和人机交互等多个领域。动作识别可以通过多种数据模态实现，包括 RGB 视频和骨骼数据。与 RGB 视频相比，骨骼数据展现出一些优势，它对身体特征的变化具有鲁棒性，并且几乎不受变化环境、复杂背景、光照条件和其他噪声源的影响。此外，深度传感器和人体姿态估

1. 广东工业大学自动化学院 广东广州 510006

计算机与数字工程 , 2023, 51(2): 440-444.

[7]TANG D W, KUPPENS P, GEURTS L, et al. End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network[J/OL].Eurasip journal on audio, speech, and music processing.2021[2025-03-26]. https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-021-00208-5.

[8]WEI S Y, ZOU S, LIAO F F, et al. A comparison on data augmentation methods based on deep learning for audio classification[J].Journal of physics: conference series, 2020, 1453: 012085.

[9]ZHAO J, ZHANG W Q.Improving automatic speech recognition performance for low-resource languages with self-supervised models[J].IEEE journal of selected topics in signal processing, 2022,16(6):1227-1241.

[10]HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J].IEEE/ACM transactions of audio, speech, and language processing,2021,29:3451-3460.

[11]YANG S L, YU Z T, WANG W J, et al. Sequence modeling[C]// Proceedings of the 23rd Chinese National Conference on Computational Linguistics. Brussels: ACL, 2024: 625-636.

[12]CHEN Z, SHAO Y F, MA Y, et al. Improving acoustic scene classification in low-resource conditions[C]//ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2024:12-30.

【作者简介】

汪慧娟（1999—），女，河南信阳人，硕士研究生，研究方向：计算机技术、移动互联应用及开发技术。

李宏滨（1968—），男，山西晋中人，硕士，副教授，研究方向：智能图像处理。

（收稿日期：2025-04-10 修回日期：2025-09-08）