持续自监督的跨语种帕金森病检测方法

石玥[1]，高倩格[1]，季薇[1]，李云[2]

（1. 南京邮电大学通信与信息工程学院，南京 210003；2. 南京邮电大学计算机学院，南京 210023）

摘　要：受到医学伦理、录制条件、受试者配合程度和标注成本等因素的影响，高质量的帕金森病语音标注数据比较稀缺，这在一定程度上制约了监督学习方法在基于语音的帕金森病检测中的应用。尽管可以通过跨语种检测和多类型语料的使用来克服上述问题，但是已有研究大多采用同质语料对齐策略，忽视了自然语境中语音表达的复杂性与多样性，难以充分提取帕金森病的深层病理特征。为此，本文提出一种持续自监督的跨语种帕金森病检测方法，以期在充分利用语音数据资源开展帕金森病检测的同时，避免费时费力的人工标注。首先，基于多类型源语种语音数据，采用顺序微调策略对自监督特征提取模型的顶层进行适应性训练。为防止语料冲突与遗忘问题，引入持续学习重演缓冲区和特征蒸馏机制，在增强自监督特征提取模型可塑性的同时提升其对旧知识的保留能力。最后，将该模型迁移至目标语种，借助自监督模型对“语种无关”特征的建模能力，实现跨语种场景下帕金森病的有效识别。在意大利语和汉语帕金森病数据集上的实验结果显示：所提方法提取的特征更具判别力，在跨语种帕金森病检测中分类精度与泛化能力均优于已有方案。

关键词：帕金森病；特征提取；跨语种；多类型语音；持续自监督学习

中图分类号：TP391　　**文献标识码：**A

A Continual Self-Supervised Method for Cross-Lingual Parkinson’s Disease Detection

SHI Yue[1]，GAO Qiange[1]，JI Wei[1]，LI Yun[2]

(1. School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China；

2. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Due to factors such as medical ethics, recording conditions, participant cooperation, and annotation costs, high-quality Parkinson's disease (PD) speech annotation data is relatively scarce, which to some extent restricts the application of supervised learning methods in speech-based PD detection. Although the above problems can be overcome through cross-lingual detection and the use of multi-types of corpora, most existing studies have adopted homogeneous corpus alignment strategies, ignoring the complexity and diversity of speech expression in natural contexts, making it difficult to fully extract the deep pathological features of PD. To this end, a continuous self-supervised cross-lingual PD detection method is proposed in this paper, aiming to fully utilize speech data resources for PD detection while avoiding time-consuming and laborious manual annotation. First, based on multi-type speech data in the source language, a sequential fine-tuning strategy is adopted to adaptively train the top-level of the self-supervised feature extraction model. To prevent corpus conflicts and forgetting issues, a continuous learning replay buffer and feature distillation mechanism are introduced to enhance the plasticity of the self-supervised feature extraction model while improving its ability to retain old knowledge. Finally, the model is transferred to the target language and the ability of self-supervised models to model "language independent" features is utilized to achieve effective recognition of PD in cross-lingual scenarios. The experimental results on Italian and Chinese PD datasets show that the proposed method extracts more discriminative features and has better classification accuracy and generalization ability than existing methods in cross-lingual PD detection.

Key words: Parkinson’s disease detection; feature extraction; cross-lingual; multi-type speech; continual self-supervised learning

引　　言

基金项目：江苏省高校基础科学（自然科学）重大项目（21KJA520003）资助 。

帕金森病(Parkinson's disease, PD)是一种常见的神经退行性疾病，早期患者常因发声系统功能失调而出现特征性语音异常，主要表现为言语流畅性下降、声强减弱、发音清晰度降低及韵律特征缺失等[1-2]。

基于语音的帕金森病辅助检测已取得了一定的效果[3]。然而，受医学伦理、录制条件、标注成本等因素制约，已公开的帕金森病语音数据集普遍存在数据规模有限、高质量标注数据不足等问题，在监督学习中容易导致模型过拟合，从而影响模型对未知样本的泛化能力[4-7]。为充分利用有限的数据资源，一些研究尝试融合包括持续元音、重复音节和阅读文本在内的多类型语音数据[8]以挖掘更全面的病理特征，如文献[8]提出一种多源语音信息融合模型，以充分挖掘持续元音与重复音节语料中的病理信息；文献[9]从多类型语音信号中提取了调音和韵律信息，通过特征解耦获得共有表示和私有表示，进而使用跨类型的分层注意力机制，以不同粒度交互的方式逐步融合共有表示和私有表示，提升了帕金森病的检测性能。但是，这些有监督的学习方法无法克服对高质量人工标注的依赖性。

此外，已公开的各数据集中受试者使用的语种各异。尽管有研究表明帕金森病导致的语音变化（如语速、发声缺陷等）具有一定的跨语种普遍性，然而基于单一语种语音数据训练的模型，往往难以在另一语种的数据上维持同样高的识别准确率[10]。当前，跨语种场景下的帕金森病检测研究还处于探索阶段，现有方法主要围绕多语种混合训练[11-14]或域自适应技术[15]展开。文献[11]提出了一个面向英语、韩语和泰米尔语的多语言帕金森病诊断框架，该方案通过建模语言特异性特征差异，显著提升了模型在各语种独立测试时的分类准确率。文献[12]结合中国台湾与韩国的语音数据，通过提取多维度的声学与语言特征，利用机器学习算法区分帕金森病患者与健康个体。文献[13]通过在捷克语、德语和西班牙语的帕金森病语音数据集上对比实验发现，单一语种训练的模型在跨语种测试时识别准确率显著降低，而采用多语种混合训练策略能有效提升模型的跨语种适应能力，且性能改善与训练数据的语种多样性呈正相关。文献[14]对比分析韩语和美式英语的帕金森病患者的语音特征，发现不同语种患者在元音共振峰分布等声学参数上存在明显区别。文献[15]提出了一种结合对抗迁移学习与特征解耦的跨语种帕金森病声学分析模型，通过提取域不变病理信息，在英语和汉语数据集上取得了良好的检测性能。然而，这些方法大多依赖于对齐的同类型语料（如只使用持续元音），忽视了自然语境中多类型语音表达的复杂性与多样性，限制了模型对复杂语音模式的捕捉能力，导致在实际应用中的鲁棒性不足。

近年来，自监督学习技术为低资源场景下的语音特征提取提供了新思路[16], 以 Wav2Vec 2.0、HuBERT（Hidden-unit BERT）等为代表的模型能够从大规模无标注语音中学习鲁棒的语音表示，并在语音识别、病症检测等任务中展现出强大的迁移能力[17]。特别地，HuBERT 模型可通过掩蔽预测隐藏单元的策略学习离散化的语音单位表示，有助于抑制语言表层差异，更突出与病理相关的运动模式变化，这一特点使其成为解决跨语种帕金森病检测问题的理想候选模型。

此外，在帕金森病语音检测任务中，不同类型语料所携带的病理特征存在显著差异，语音特征的分布也因语料类型而异。若将多种语料混合训练，模型在学习新语料特征时容易覆盖甚至遗忘先前学习的语料特征，造成数据冲突与灾难性遗忘[18]，严重影响模型的稳定性与泛化能力。持续学习[19-20]是解决这一问题的有效途径，基于重演策略的持续学习是其中一种比较主流的方法。该方法通过显式保留部分先前任务的数据或中间特征表示，使保留的数据或特征在训练新任务时与当前数据或特征共同参与学习，从而缓解灾难性遗忘。其中，基于样本的重演方法，如增量式分类与表示学习[21]将代表性样本存储于有限的缓冲区中进行周期性回放；而基于特征重演的方法，如贪婪采样与简单学习器[22]则保留先前样本的特征或生成表示，通过模型自身内部机制进行重构与更新。整体而言，重演方法在任务间分布差异较大的情况下，能够显著提升模型的稳定性与持续学习能力。

虽然，已有的跨语种帕金森病检测方法可以通过对齐同类型语料以缓解语种迁移困难，但语料类型的单一化限制了模型对复杂语音模式的捕捉能力，导致在实际应用中的鲁棒性不足。因此，迫切需要一种无需大量标签、能同时适应多语种与多语料环境的鲁棒帕金森病检测方法。为此，本文提出一种持续自监督的跨语种帕金森病检测方法，以期在充分利用语音数据资源开展帕金森病检测的同时，避免费时费力的人工标注。本文的主要贡献在于：

（1）基于多类型帕金森病语音数据的持续自监督学习框架：为充分利用多类型帕金森病语音数据资源，设计了一种阶段式顺序微调策略对 HuBERT 自监督特征提取模型的顶层进行适应性训练，使其能依次适应不同类型的语音数据，并逐步引导其提取更深层次的、与帕金森病检测任务相关的判别性特征。

（2）有针对性的抗遗忘机制：为防止语料冲突与遗忘问题，引入了结合 K-means 聚类[23]采样的重演缓冲区和基于余弦相似度的特征蒸馏（Feature distillation, FD）机制，有效缓解了模型在连续学习多种语料时的灾难性遗忘问题，平衡了对旧知识的保留能力和对新知识的学习能力。

（3）语种无关的病理特征迁移：得益于自监督模型对深层、语种无关特征的强大建模能力，将在源语种上训练好的模型直接迁移至目标语种进行特征提取和分类，实现了无需目标语种标注数据的有效跨语种检测。

在意大利语和汉语帕金森病数据集上的实验结果表明，本文所提方法能够提取更具判别力的帕金森病语音特征，在跨语种帕金森病检测中的分类精度与泛化能力均优于已有方案。

1　基于持续自监督跨语种语音特征提取的帕金森病检测方法

本文提出一种基于持续自监督跨语种语音特征提取的帕金森病检测模型，以缓解高质量标注数据稀缺问题，提升语种间的迁移能力。模型总体可分为两部分：源语种上的持续自监督帕金森病语音特征提取和目标语种上的语音特征提取和帕金森病检测，整体框架如图 1 所示。

在源语种上，选取包含多种语料类型（如持续元音、重复音节和阅读文本）的帕金森病语音数据库作为源域数据集，用于自监督特征提取模型的训练。为了充分挖掘语音中的深层特征，本文加载预训练好的 HuBERT 模型作为初始化自监督特征提取模型H_0。每次训练仅使用一种类型的语料，并采用顺序微调策略对自监督 HuBERT 模型的顶层进行适应性训练，以逐步引导其提取更深层次、与帕金森病检测任务相关的判别性特征。在此基础上，为缓解模型在新语料训练过程中遗忘旧语料知识的问题，引入持续学习机制，结合重演策略和特征蒸馏保留历史知识[24]。

在目标语种上，使用微调过的 HuBERT 自监督特征提取模型，对目标语种语音进行特征提取。提取后的特征输入至训练好的分类器中，实现对目标语种中帕金森病患者的有效识别。

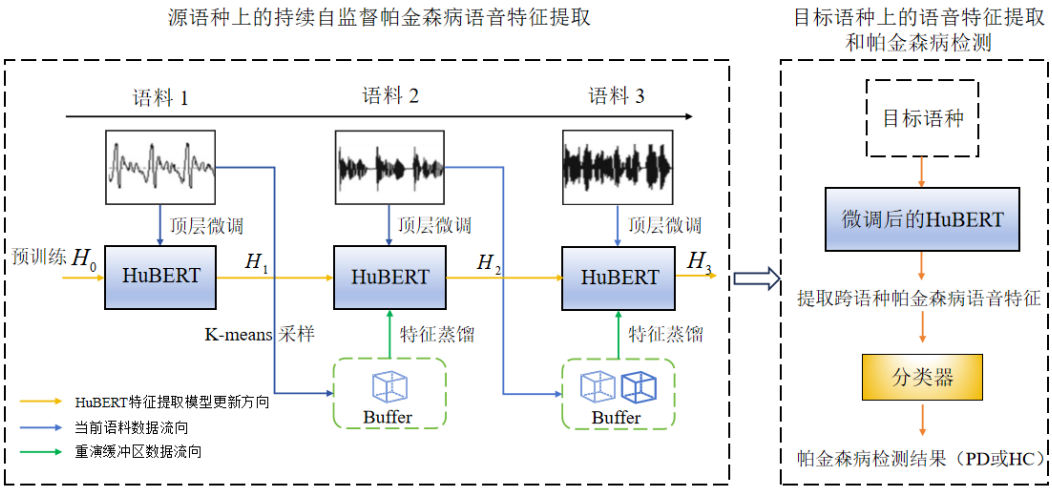

图 1　持续自监督的跨语种帕金森病检测模型

Fig.1 Continuous self-supervised cross-lingual Parkinson's disease detection model

1.1　自监督特征提取模型

本研究使用 HuBERT 模型作为帕金森病语音特征提取模型，如图 2 所示，该模型由卷积神经网络（Convolutional neural network, CNN）编码器、Transformer 编码器以及伪标签生成模块构成。这里，HuBERT 不直接使用原始语音波形作为监督信号，而是采用一个两阶段的伪标签生成策略：

（1）掩蔽预测阶段

在帕金森病语音检测任务中，患者常表现为发音不清、单调语调及讲话断续等全局性时序异常特征。HuBERT 采用块状掩蔽策略 (Block-wise masking)[25]，在语音帧序列上随机选择起始位置，在连续时间步上掩蔽 15%的帧，使得模型在预测被遮蔽帧的聚类标签时，必须综合利用掩蔽段前后更大范围的上下文信息，包括患者语速变化、音调波动及

共振峰迁移等跨越多帧的全局性信号模式，从而学习到能够表征帕金森病患者全局语音异常的高阶表示。图 3 展示了 HuBERT 块状掩蔽策略应用于 Mel 语谱图后的可视化效果。需要说明的是，15%的比例是自监督语音学习领域经过广泛验证的常用设置，能在提供足够上下文线索和保证任务难度之间取得良好平衡。

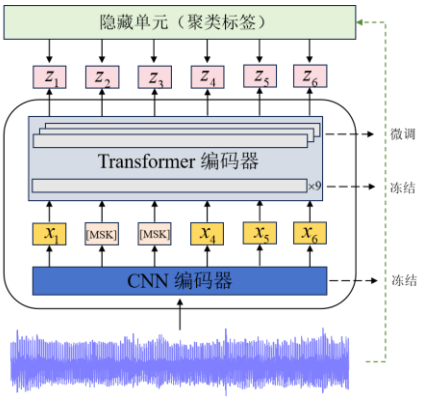

图 2 HuBERT 自监督语音特征提取模型结构

Fig.2 Architecture of the HuBERT self-supervised speech feature extraction model

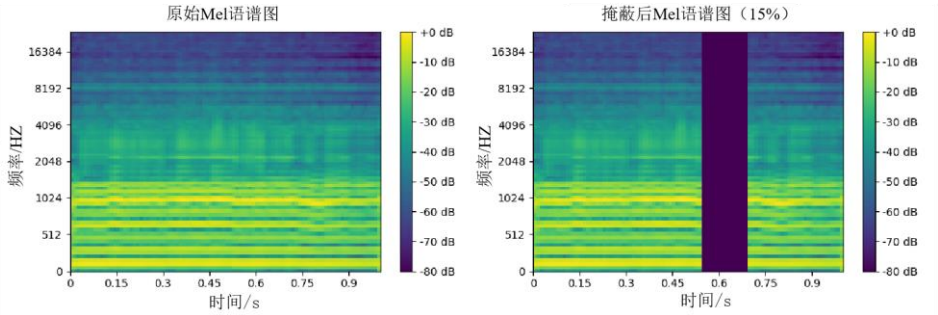

图 3 以 Mel 语谱图为例的块状掩蔽策略

Fig.3 Block masking strategy with Mel spectrogram as an example

具体地，首先从原始语音信号计算梅尔倒谱系数(Mel frequency cepstral coefficient, MFCC)特征，并使用 K-means 聚类方法将其聚类成 K 个类别。使用 CNN 编码器将原始音频波形转换为帧级特征$x_i, 1 \le i \le n$，n 为帧数，每一帧都会被分配一个聚类标签，作为该时间帧的目标伪标签。根据前述的块状掩蔽策略对帧级特征进行掩蔽，掩蔽后的特征将作为 Transformer 的输入，由 Transformer 执行掩蔽预测任务。HuBERT 的目标是预测被掩蔽帧的聚类标签，而不是直接预测原始波形或特征。因此，Transformer 的输出是一个时间序列向量$\boldsymbol{Z} = (z_1, z_2, \cdots, z_i, \cdots z_n)$，映射到 K 维分类概率分布，第 i 个帧输出为$\boldsymbol{P}_i = [p_1, p_2, \cdots, p_k, \cdots, p_K]$，其中，$k$ 为聚类标签序号，p_k则是对应聚类标签的预测概率。这里，采用负对数似然损失来执行掩蔽预测任务，通过反向传播优化 Transformer 的参数，使其能够更准确地预测被掩蔽的帧，即

$$L_h = -\sum_{i \in \boldsymbol{M}} \log_2 \boldsymbol{P}_i (C_i \mid \boldsymbol{X}_{unmasked}) \tag{1}$$

式中：$\boldsymbol{M}$ 是掩蔽的时间帧集合，C_i 是第 i 帧的聚类标签，$\boldsymbol{X}_{unmasked}$ 是未被掩蔽的帧。

（2）自监督学习阶段

在该阶段，HuBERT 不再使用 MFCC 特征聚类生成的聚类标签作为监督信号，而是利用 K-means 对 Transformer 中间层的特征进行聚类，生产更准确的聚类标签，再使用新的聚类标签对模型继续训练。考虑到帕金森病语音数据在资源上存在稀缺性，直接对整个 HuBERT 模型进行全参数微调可能导致过拟合，并破坏原始模型在预训练阶段所学到的通用语音特征。为充分利用预训练模型的表征能力，同时降低对小样本数据的依赖，本文仅对 HuBERT 模型的顶层 Transformer 模块进行微调。具体而言，微调过程中冻结 HuBERT 的所有底层参数，仅解冻顶层 Transformer 最后 3 层，并且每个阶段在单一类型语料上进行训练。通过这种方式，保留底层已学习的语音基础表示，仅在高层特征空间中引

入任务相关的语义信息，从而使模型逐步适应帕金森病检测任务。

1.2 阶段式顺序训练

为避免帕金森病语音检测任务中多种语料混合训练导致的数据冲突与灾难性遗忘，本文引入按序输入策略，在模型训练阶段逐类输入不同语料，以降低语料间的特征干扰，提升模型在多语料环境下的适应性。

阶段式顺序训练的具体实现如图 1 中的源语种自监督特征提取模型微调模块所示。首先，取包含多种语料类型的帕金森病语音数据库作为源语种数据集，包括：持续元音数据集、重复音节数据集与阅读文本数据集。加载预训练好的 HuBERT 自监督模型作为源语种微调训练的初始化模型H_0。训练时，模型采用分阶段训练方式：每阶段仅使用一种类型的语料对 HuBERT 模型的顶层进行微调，并结合 Adam 优化器对顶层三层进行参数更新。

为应对分阶段训练中可能引发的灾难性遗忘问题，本文引入持续学习机制，并采用重演策略辅助模型在新阶段训练中保持对旧任务的记忆。具体地，在每一阶段训练完成后，当前模型参数被保留并作为下一阶段训练的初始状态。随后，引入 K-means 聚类算法，从当前阶段所使用的语料$\boldsymbol{D}_t, t = 1,2,3$中选取具有代表性的样本，构建重演缓冲区$B$，以缓解后续阶段可能引发的灾难性遗忘问题。

1.2.1 重演缓冲区管理

为解决多类型语料训练的灾难性遗忘问题，重演缓冲区的构建与数据增强均在 HuBERT 模型提取的高层特征域执行。本文采用 K-means 采样策略对各语料进行样本选择，以替代在捕获数据多样性方面不具鲁棒性的随机抽样方法。首先，利用当前阶段微调后的 HuBERT 模型$H_t, t = 1,2,3$，提取所有语料的高层特征（Transformer 编码器最后一层输出的全局平均池化特征），并对该特征域数据进行 K-means 聚类，划分为 C 个簇。聚类簇数 C 并非直接指定，而是通过目标缓冲区大小反向确定，为了使采样过程覆盖数据分布的主要模式，我们设定聚类簇数 C 等于缓冲区大小的一半，以确保每个聚类有足够样本供后续选择并捕捉更细致的分布特征。完成聚类后，从每个簇中分别选取最接近聚类中心的样本和分布边缘的样本，以兼顾典型性与多样性，从而更有效地构建具有代表性的重演缓冲区。最后，将所有 C 个簇选出的 $2C$ 个样本存入重演缓冲区。缓冲区中存储的是聚类后选中样本的高层特征向量及对应原始音频索引。

为增加重演缓冲区的多样性，防止模型过拟合，可在特征域对缓冲区样本进行二元混合(Instance mixup, IM)数据增强，生成混合特征向量：$x' = \lambda x_1 + (1-\lambda)x_2$，其中，$\lambda$ 是[0,1]之间的均匀随机变量；x' 是增强后的特征向量；x_1、x_2 分别是从重演缓冲区B中随机选取的高层特征向量。

考虑到帕金森病语音数据集普遍容量较小且各类型语料的语音特征分布存在显著差异，为增强模型对多语料任务的适应性与鲁棒性，本文在所提方法中采用累积保留所有阶段代表性样本的重演策略，每阶段保留 10%的当前类型语料数据，使各语料类型携带的病理语音线索形成互补，进而助力构建对目标语种更全面、更细粒度的判别能力。

1.2.2 特征蒸馏

在随后的训练阶段，模型在新语料上继续对自监督特征提取模型进行训练，同时利用重演缓冲区中存储的高层特征向量，分别与H_{t-1}、H_t模型输出的特征向量进行对齐，通过特征蒸馏约束模型参数更新，并反向传播更新当前自监督特征提取模型H_t，详见图 4。该蒸馏过程将引导当前阶段的自监督特征提取模型保留对先前语料的判别能力，并在此基础上对H_t参数进行进一步优化。

本文采用的特征蒸馏旨在对齐帕金森病相关语音表征以保留历史知识，而非用于模型压缩，因此不涉及模型复杂度。若采用L_1损失，损失如下所示

$$L_1 = \sum_{i=1}^{n} | \boldsymbol{Z}_{i,t} - \boldsymbol{Z}_{i,t-1} | \tag{2}$$

式中 $\boldsymbol{Z}_{i,t}$ 和 $\boldsymbol{Z}_{i,t-1}$ 分别为自监督模型H_t和H_{t-1}的第 i 个样本的输出特征，其维度与 HuBERT 模型的隐藏层维度保持一致（768 维），且与后续分类器的输入维度相互独立。L_1损失强制两个特征对齐，即使两个特征最后的预测结果一致，只要特征存在差异，就会增加损失，进而影响模型的更新方向，导致模型过拟合于旧模型参数，对于新语料的判别能力下降。

本文希望自监督特征提取模型保留旧知识的同时，允许它在新语料在学习到新知识。为此，这里采用余弦相似度作为正则项进行特征蒸馏，余弦相似度是衡量两个向量方向一致性的经典指标[26]，其核心特性是仅关注向量间的方向

关系，不约束向量幅度，恰好契合本文“保留旧知识同时兼容新知识”的需求。基于余弦相似度的特征蒸馏损失L_f定义为所有样本上余弦相似度的负和，旨在最大化新旧模型输出特征方向的一致性，即

$$L_f = -\sum_{i=1}^{n} S(\boldsymbol{Z}_{i,t}, \boldsymbol{Z}_{i,t-1}) \tag{3}$$

式中$S(\boldsymbol{Z}_{i,t}, \boldsymbol{Z}_{i,t-1})$用于计算两个自监督模型$H_t$和$H_{t-1}$的输出特征的余弦相似度。

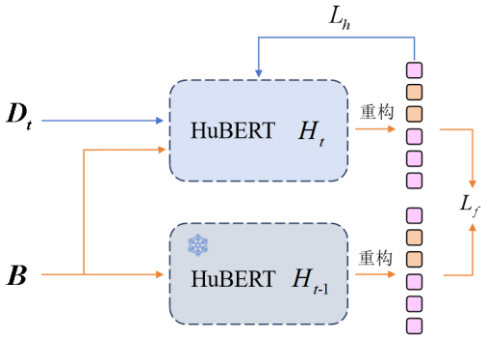

图 4　特征蒸馏过程

Fig.4 Feature distillation process

1.2.3 阶段式顺序训练执行流程

所提方法的阶段式顺序训练执行流程如算法1所示。

算法 1：所提方法的阶段式顺序训练流程

输入：源语种三种不同语料数据$\{\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3\}$和预训练好的 HuBERT 特征提取模型$H_0$。
输出：适配跨语种帕金森病检测任务的自监督特征提取模型H_3。
阶段 1：
　　在$\boldsymbol{D}_1$上微调H_0，计算掩蔽预测损失并反向传播更新模型参数，得到H_1。
　　K-means 聚类采样构建重演缓冲区$\boldsymbol{B} \leftarrow Sample(\boldsymbol{D}_1)$并对缓冲区进行数据增强。
阶段 2（t=2）和阶段 3（t=3）：
　　训练数据$\boldsymbol{D} \leftarrow \boldsymbol{D}_t \cup \boldsymbol{B}$
　　for epoch = 1, 2, …:
　　　遍历训练数据集的所有批次，依次读取每个批次的样本 x
　　　if $x \in \boldsymbol{D}_t$:
　　　　进行掩蔽预测任务，计算损失反向传播更新自监督特征提取模型H_t。
　　　else $x \in \boldsymbol{B}$:
　　　　对H_{t-1}和H_t的输出进行特征蒸馏，计算蒸馏损失并反向传播更新自监督特征提取模型H_t。
　　　end if
　　end for
　　$\boldsymbol{B} \leftarrow \boldsymbol{B} \cup Sample(\boldsymbol{D}_t)$

1.3　目标语种上的语音特征提取和帕金森病检测

经过基于源语种多类型语音数据对自监督特征提取模型的阶段式训练后，在目标语种数据集上，加载在源域数据上微调过的自监督特征提取模型H_3，对目标语种语音进行深层病理特征提取。得益于自监督模型对语音信号中“语言无关”结构性特征的建模能力，其提取的高层表示具有良好的迁移性与泛化性，能够跨越语种差异捕捉帕金森病相关的语音异常。提取后的特征经过进一步处理和降维后输入至分类器中，实现对目标语种中各类型语音帕金森病的有效检测。这里，分类器是一个参数独立的两层全连接网络(维度[16, 2])。

2　实验

2.1　数据集简介

意大利语帕金森病语音数据集[27]由巴里大学的一支研究团队创建，该数据集收集了 65 名受试者的 3 种类型的帕金森病语音数据：持续元音(/a/、/e/、/i/、/o/、/u/)、重复音节(/pa-ta/)以及指定文本和单词的朗读。其中，用于测试的文本是语音平衡的，受试者以距离麦克风 15 cm 到 25 cm 的距离朗读文本。受试者的语音经过编辑，最终得到总计

650 个元音样本、130 个重复音节样本和 100 个阅读指定文本样本，所有样本均以.wav 格式保存。意大利语帕金森病语音数据集的统计信息详见表 1。

第二个语音数据集为本研究团队与南京医科大学附属老年医院神经内科合作创建的自采汉语帕金森病语音数据集。受试者由该医院帕金森病及运动障碍专病门诊筛选出的 68 名 PD 患者和 17 名健康人组成的健康对照组(Healthy controls, HC)的语音数据构成，数据采集均在安静的室内环境进行。受试人群的统计信息详见表 2，其中，HY 分级表示对帕金森病变程度的评估，3 期以前属于轻中度，3 期及以上患者症状显著加重。采集内容为持续元音/a/，重复音节(/pa-ka-la/)以及指定文本和单词的朗读，经剪辑处理后共计得到 270 个元音样本、170 个重复音节样本以及 60 个指阅读指定文本样本。

表 1 意大利语帕金森病数据集的统计信息

Table 1 Statistical information of the Italian Parkinson's disease dataset

	男性		女性		合计	
受试者类别	PD	HC	PD	HC	PD	HC
受试者人数	19	23	9	14	28	37
平均年龄及统计方差	67.2 (8.3)	47.6(23.7)	67.2(9.3)	48(23.4)	67.2(8.3)	48.3(23.4)
年龄分布	50~80	19~77	40~80	19~69	40~80	19~77

表 2 自采的汉语帕金森病语音数据集的统计信息

Table 2 Statistical information of the self-collected Chinese Parkinson's disease speech dataset

	男性		女性		合计	
受试者类别	PD	HC	PD	HC	PD	HC
受试者人数	49	8	19	9	68	17
平均年龄及统计方差	69.3(9.5)	66.5(7.2)	69.8(8.2)	65.3(6.8)	69.4(9.2)	65.9(7.0)
年龄分布	46~88	58~77	56~84	53~74	46~88	53~77
平均病情持续时间及统计方差	5.9 (3.6)	0	5.4 (3.1)	0	5.8 (3.4)	0
HY 分期	1~4	0	1~4	0	1~4	0

2.2 实验设置

所有实验均在 2.1 节所述的两种多类型语种数据集上进行，模型的实验参数设置如表 3 所示。其中，Dropout (p=0.05) 应用于分类器部分的两层全连接网络之间，旨在缓解小规模数据集上可能出现的过拟合问题，提升模型的泛化能力。

为了充分验证本文所提方法的有效性，选择准确率(Accuracy, ACC)、敏感度(True positive rate, TPR)、特异度(True negative rate, TNR)作为性能评价指标，其计算公式分别为

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \tag{4}$$

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$TNR = \frac{TN}{TN + FP} \tag{6}$$

式中：TP（True positive）表示正确识别出的帕金森病患者数量，TN（True negative）表示正确识别出的健康个体数量，FP（False positive）表示将健康个体误判为患者的数量，FN（False negative）表示将帕金森病患者漏判的健康个体数量。

表 3 模型参数信息

Table 3 Model parameter information

参数	参数值
分类器	[16,2]
重演缓冲区采样率（%）	10
batch size	32
epoch	60
学习率	0.000 1
Dropout	0.05

优化器	Adam

2.3 实验结果

2.3.1 自监督特征提取模型有效性验证

为验证本文所使用的自监督特征提取模型 HuBERT 的有效性，本节对不同特征提取模型在跨语种帕金森病检测任务中的表现进行了比较。为保证实验条件的一致性，所有模型均采用相同的阶段式顺序进行训练，利用源域数据对自监督模型进行微调，在目标语种上进行帕金森病检测。参与比较的模型包括 CNN+MFCC、Transformer[28]和 Wav2Vec 2.0[29]。其中，“CNN+MFCC”使用 CNN 作为特征提取器，将 MFCC 特征输入 CNN 中提取高级特征。

为公平比较不同模型在跨语种帕金森病检测任务上的性能，所有模型采用统一的分层渐进式微调策略，旨在控制变量，使性能差异更可能归因于模型架构与预训练范式本身，而非微调程度的差异。核心是在微调过程中，仅解冻并更新模型的最后若干层参数，同时冻结其余所有底层参数。HuBERT 与 Wav2Vec 2.0 模型均为大规模预训练的自监督模型，参数量巨大，为了保留它们在预训练阶段学到的通用语音表征能力且避免在小规模帕金森病数据集上过拟合，我们遵循其原始论文及通用实践，仅微调 HuBERT 模型最后 3 层 Transformer 模块和 Wav2Vec 2.0 模型最后 4 层 Transformer 模块（因其上下文网络总层数通常为 12 或 24 层，微调顶部 4 层是常见设置），底层 CNN 特征提取器及其他 Transformer 层均被冻结。由于 Transformer 基准模型并非从头训练，而是加载在大型语料库上预训练好的权重，因此为保证对比的一致性，仅微调其顶部 2 层 Transformer 编码器，冻结其余层。CNN+MFCC 基准模型不涉及预训练，其 CNN 部分为随机初始化，为保证其能充分学习任务特征，在训练时对其所有参数进行整体更新，这对于参数较少的传统模型是公平且必要的。上述策略可有效避免因微调参数量级不同而带来的不公平性，使性能差异更能归因于模型本身的特征提取与泛化能力。

同时，我们充分尊重不同模型的原生学习范式，为其采用最合适的微调损失函数。HuBERT 模型继续使用其固有的掩蔽预测损失，通过预测被掩蔽帧的伪标签来学习上下文相关的特征。Wav2Vec 2.0 模型采用其标准的对比学习损失，使其在区分真实上下文表示与干扰项的过程中优化特征。Transformer 与 CNN+MFCC 基准模型采用适用于分类任务的交叉熵损失进行端到端微调。本文认为，这种“因模制宜”的微调策略不仅公平，而且是科学对比的必要条件。

以汉语为源语种，意大利语为目标语种，基于多类型语音数据开展的跨语种帕金森病检测结果如表 4 所示。以意大利语为源语种，汉语为目标语种，基于多类型语音数据开展的跨语种帕金森病检测结果如表 5 所示。实验中，自监督特征提取模型在源域训练时，输入语料的顺序均为：持续元音、重复音节、阅读文本。为探索目标语种语音数据类型的不同对检测结果的影响，表 4 和表 5 均根据目标语种的语音数据类型不同，对跨语种帕金森病检测的结果进行了分别展示。从表 4 和表 5 的结果可以发现，本文所使用的自监督特征提取模型 HuBERT 在跨语种帕金森病检测任务中均取得了较好的性能。实验结果表明，HuBERT 所采用的掩蔽预测范式在本研究的跨语种场景下展现出独特优势，分类准确率相较于 Wav2Vec 2.0 模型提升了约 0.9-1.0 个百分点，证明 HuBERT 学习离散语音单元的特性，能有效抑制语种特异性干扰，更专注于提取跨语种共享的帕金森病病理特征。就本文方法而言，当目标语种中使用持续元音类型的语音数据时，帕金森病的检测效果最佳，这是因为与其他两种类型的语音数据相比，持续元音更具主导作用，这一结论在后续的实验中均有体现。

表 4 不同自监督特征提取模型在跨语种帕金森病检测任务中的性能比较（汉语→意大利语）

Table 4 Performance comparison of different self-supervised feature extraction models on cross-lingual Parkinson's disease detection task (Chinese → Italian) %

自监督模型	目标语种下的语音数据类型								
	持续元音			重复音节			阅读文本		
	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR
CNN+MFCC	89.0	87.6	89.4	88.8	88.5	89.3	88.3	87.9	88.6
Transformer	89.6	88.7	90.2	89.2	89.0	89.7	88.5	88.2	88.9
Wav2Vec 2.0	90.8	91.1	90.4	89.7	89.1	90.4	88.9	88.0	**89.5**

Ours(HuBERT)	**91.7**	**91.4**	**92.2**	**90.6**	**89.7**	**91.1**	**88.9**	**88.5**	89.3

表 5 不同自监督特征提取模型在跨语种帕金森病检测任务中的性能比较（意大利语→汉语）

Table 5 Performance comparison of different self-supervised feature extraction models on cross-lingual Parkinson's disease detection task (Italian → Chinese) %

自监督模型	目标语种下的语音数据类型								
	持续元音			重复音节			阅读文本		
	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR
CNN+MFCC	88.6	88.2	88.9	88.1	87.8	88.4	87.7	87.3	87.9
Transformer	89.1	88.7	89.3	88.5	88.3	88.8	87.9	87.7	88.3
Wav2Vec 2.0	90.3	90.4	90.7	89.4	89.1	89.7	88.2	87.7	88.6
Ours(HuBERT)	**91.3**	**90.9**	**91.6**	**90.3**	**89.5**	**90.8**	**88.4**	**88.0**	**89.3**

2.3.2 持续学习方法的性能比较

为验证本文所提持续学习方法的有效性，本节将不同持续学习方法在跨语种帕金森病检测任务上的性能进行了比较。参与比较的方法对重演样本的选择方式（随机采样、聚类采样）和特征蒸馏损失（L_1损失、基于余弦相似度构建的L_f损失）进行了不同组合。

以汉语为源语种，意大利语为目标语种，不同持续学习方法在跨语种帕金森病检测任务上的性能如表 6 所示，自监督特征提取模型的语料输入顺序同 2.3.1 节。从表 6 的结果可知，本文所提的使用 K-means 聚类构建重演缓冲区并结合余弦相似度进行特征蒸馏的方案取得了最优的检测性能。这是由于随机采样缺乏明确的结构信息，容易引入噪声或不具判别性的样本，而 K-means 重演的方式能更好地引导模型关注帕金森病语音中的核心发音单位，提升了模型在识别帕金森病相关语音特征时的稳定性和准确性。采用L_1损失作为正则项会导致模型可塑性降低，难以有效学习新特征，而基于余弦相似度构建的损失函数可在不遗忘旧知识的同时，保证模型对新特征的可塑性。

表 6 不同持续学习方法在跨语种帕金森病检测任务中的性能比较（汉语→意大利语）

Table 6 Performance comparison of different continual learning methods on cross-lingual Parkinson's disease detection task (Chinese → Italian) %

持续学习方法	目标语种下的语音数据类型								
	持续元音			重复音节			阅读文本		
	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR
随机采样+ L_f损失	88.4	88.1	88.9	88.2	87.9	88.5	86.9	87.1	86.6
随机采样+ L_1损失	89.9	89.4	90.3	89.5	89.0	89.7	87.7	87.3	88.5
仅中心+L_1损失	88.5	88.1	89.0	89.2	88.7	89.9	87.7	87.0	88.3
仅中心+L_f损失	90.5	90.1	90.9	89.4	88.9	90.1	87.8	87.3	88.4
聚类采样+ L_1损失	88.6	88.1	89.3	88.3	88.1	88.6	87.3	87.1	87.7
聚类采样+ L_f损失	**91.7**	**91.4**	**92.2**	**90.6**	**89.7**	**91.1**	**88.9**	**88.5**	**89.3**

2.3.3 阶段式训练顺序分析

为探究阶段式训练中不同语料的呈现顺序对跨语种帕金森病检测性能的影响，本节对源域语种中的 3 类语料进行了全排列组合实验。实验共包含 6 种不同的语料顺序排列方式，如“vpr”表示按照持续元音、重复音节、阅读文本的顺序输入，相关实验结果如表 7 所示。

表 7 不同语料训练顺序组合在跨语种帕金森病检测任务中的性能比较（汉语→意大利语）

Table 7 Performance comparison of different training sequence combinations of multilingual corpora on cross-lingual Parkinson's disease detection task (Chinese → Italian) %

源域语种语料训练顺序	目标语种下的语音数据类型								
	持续元音（v）			重复音节（p）			阅读文本（r）		
	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR
vpr	91.7	91.4	92.2	90.6	89.7	91.1	88.9	88.5	89.3
vrp	91.7	**91.5**	92.2	90.6	**89.8**	91.4	88.2	87.9	88.5
pvr	91.4	91.2	91.7	90.3	89.5	90.8	**89.0**	**88.6**	**89.3**
prv	**91.8**	91.1	**92.4**	90.4	89.7	90.4	88.6	88.3	88.9
rvp	91.5	91.2	91.8	**90.8**	89.7	**91.6**	88.5	88.3	88.8
rpv	91.5	91.3	91.9	90.5	89.6	91.2	88.6	88.4	89.0

实验结果表明，源语种上训练自监督特征提取模型时，不管语料输入顺序如何，均能在目标语种帕金森病检测任务中取得较为优异的性能。值得关注的是，当目标语种语料类型与训练过程中最后输入的语料类型相一致时，模型将表现出更佳的检测准确率。

2.3.4 重演缓冲区容量分析

在监督学习中，重演缓冲区容量越大，模型发生遗忘的可能性通常越小。然而，对于包含多种类型语料的任务而言，较大的缓冲区可能引发数据间冲突的加剧，并显著增加计算开销，从而对当前类型语料的表示学习造成干扰[30]。为在缓解遗忘与提升当前任务学习能力之间寻求最优平衡，本节在源域为自采汉语数据的 3 类语料中分别设置了 1%、5%、10%、15%和 20%的缓冲区采样率进行实验，例如每次训练从训练数据中采样 10%的数据作为缓冲区样本，目标语种使用意大利语的元音数据，其结果如表 8 所示。当缓冲区采样率从 1%增加至 10%时，模型性能显著提升；然而，当采样率进一步从 10%增加至 20%时，模型性能的提升变得微弱。考虑到帕金森病语音数据本身数据容量小，且较大的缓冲区容量可能带来过拟合风险，本文最终选择 10%的缓冲区采样率作为最佳设置，以平衡性能提升与过拟合风险，确保模型在帕金森病检测任务中的泛化能力。

表 8 不同缓冲区容量对多类型语料持续学习性能的影响

Table 8 Impact of different buffer capacities on continual learning performance with multi-type corpora %

缓冲区采样率	目标语种下的语音数据类型		
	持续元音		
	ACC	TPR	TNR
1%	89.7	89.3	90.6
5%	90.4	89.9	90.8
10%	91.7	91.4	92.2
15%	91.7	91.4	92.0
20%	**91.8**	**91.5**	**92.2**

2.3.5 消融实验

本节在源域为自采汉语数据，目标域为意大利语数据集的设置下，通过消融实验来检测所提方法中关键模块的作用，重点探索重演缓冲区数据增强以及特征蒸馏对模型的贡献，其中，w/o IM 代表重演缓冲区不进行数据增强，w/o FD 代表模型不进行特征蒸馏，w/o IM and FD 则代表数据增强和特征蒸馏功能均不具备，实验的详细结果如下表 9 所示。从实验结果可以看出，本文所提的各种策略均对模型性能的提升产生了积极影响。其中，采用余弦相似度作为正则项进行特征蒸馏显著提高了对旧知识的保留能力，并有效维持了模型对新知识的可塑性，确保了模型在跨语种任务中对帕金森病症状的识别准确性。尽管对缓冲区样本进行的数据增强在整体性能提升方面作用有限，但这一策略通过增加重演样本的多样性，有助于避免模型在旧语种数据上过拟合。

表 9 消融实验

Table 9 Ablation study %

模型	目标语种下的语音数据类型								
	持续元音			重复音节			阅读文本		
	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR
Ours（w/o IM）	91.5	91.0	91.7	90.1	89.7	90.6	88.2	88.0	88.5
Ours（w/o FD）	90.7	90.6	91.2	89.4	89.2	89.7	87.7	87.4	88.1
Ours （w/o IM and FD）	89.3	88.6	89.7	88.8	88.6	89.1	87.2	86.9	87.7
Ours	**91.7**	**91.4**	**92.2**	**90.6**	**89.7**	**91.1**	**88.9**	**88.5**	**89.3**

3 结束语

本文针对帕金森病语音检测中高质量标注数据稀缺与跨语种泛化的难题，提出了一种持续自监督的跨语种检测方法。通过设计阶段式顺序微调策略，并引入基于聚类采样的重演缓冲区与基于余弦相似度的特征蒸馏机制，克服了传统同质语料对齐方法的局限，实现了对多类型语音病理特征的持续学习与有效融合。本文所提方法在汉-意跨语种检测中取得了最高 91.7%的准确率，显著优于其他对比模型。本文对数据集的划分仅考虑了语料类型，然而帕金森病患者的语音还受到性别、年龄、患病时长以及病情严重程度等多种因素的影响。未来的研究可进一步引入这些先验信息，以增强模型对个体差异的建模能力，从而提升检测的准确性与鲁棒性。

参 考 文 献：

[1] Cuong Q N, Abdul M M, Daniel N P, et al. Computerized analysis of speech and voice for Parkinson's disease: A systematic review[J]. Computer Methods and Programs in Biomedicine, 2022, 226：107133.

[2] Morris H R, Spillantini M G, Sue C M, *et al*. The pathogenesis of Parkinson's disease[J]. The Lancet, 2024, 403(10423): 293-304.

[3] Sakar B E, Isenkul M E, Sakar C O, *et al*. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings[J]. IEEE Journal of Biomedical and Health Informatics, 2013, 17(4): 828-834.

[4] Little M, McSharry P, Hunter E, et al. Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease[J]. Nature Precedings, 2008: 1-1.

[5] Tsanas A, Little M A, McSharry P E, et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease[J]. IEEE Transactions on Biomedical Engineering, 2012, 59(5): 1264-1271.

[6] Ma J, Zhang Y, Li Y, et al. Deep dual-side learning ensemble model for Parkinson speech recognition[J]. Biomedical Signal Processing and Control, 2021, 69: 102849.

[7] Liu Y, Reddy M K, Penttilä N, et al. Automatic assessment of Parkinson's disease using speech representations of phonation and articulation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 242-255.

[8] 季薇，王传瑜，李云，等. 基于多源语音信息融合的帕金森病辅助检测方法[J]. 信号处理, 2023, 39(12): 2254-2264.
Ji Wei, Wang Chuanyu, Li Yun, *et al*. Auxiliary detection method of Parkinson’s disease based on multi-source speech information fusion[J]. Journal of Signal Processing, 2023, 39(12): 2254-2264.

[9] 吴迪，季薇，郑慧芬，等. 基于多类型语音信息分层融合的帕金森病检测模型[J]. 模式识别与人工智能，2024, 37(9): 811-823.
Wu Di, Ji Wei, Zheng Huifen, *et al*. Parkinson’s disease detection model based on hierarchical fusion of multi-type speech information[J]. Pattern Recognition and Artificial Intelligence, 2024, 37(9): 811-823.

[10] Lim W S, Chiu S I, Peng P L, et al. A cross-language speech model for detection of Parkinson’s disease[J]. Journal of Neural Transmission, 2025, 132(4): 579-590.

[11] Yeo E J, Choi K, Kim S, et al. Cross-lingual dysarthria severity classification for English, Korean, and Tamil[C]//2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). [S.l.]: IEEE, 2022: 566-574.

[12] Klempíř O, Krupička R. Analyzing Wav2Vec 1.0 embeddings for cross-database Parkinson’s disease detection and speech features extraction[J]. Sensors. 2024，24(17): 5520.

[13] Hernandez A, Pérez-Toro P A, Nöth E, et al. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition[J]. arxiv preprint arxiv: 2204.01670, 2022.

[14] Kim Y, Choi Y. A cross-language study of acoustic predictors of speech intelligibility in individuals with Parkinson's disease[J]. Journal of Speech, Language, and Hearing Research, 2017, 60(9): 2506-2518.

[15] 季薇，王传瑜，吴迪，等. 基于跨语种声学分析的帕金森病检测方法［J］. 电子与信息学报, 2024, 46(2): 546-554.

Ji Wei, Wang Chuanyu, Wu Di, *et al*. Parkinson's disease detection method based on cross-language acoustic analysis [J]. Journal of Electronics & Information Technology, 2024, 46(2): 546-554.

[16] 季薇，杨茗淇，李云, 等. 基于掩蔽自监督语音特征提取的帕金森病检测方法[J]. 电子与信息学报, 2023, 45(10): 3502-3510.
Ji Wei, Yang Mingqi, Li Yun, *et al*. Parkinson's disease detection method based on masked self-supervised speech feature extraction[J]. Journal of Electronics & Information Technology, 2023, 45(10): 3502-3510.

[17] 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 on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[18] Cho S, Lee H, Baek S, et al. Neuromimetic metaplasticity for adaptive continual learning without catastrophic forgetting[J]. Neural Networks: the Official Journal of the International Neural Network Society, 2025, 190: 107762.

[19] Hou S, Pan X, Loy C C, et al. Lifelong learning via progressive distillation and retrospection[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 437-452.

[20] Parisi G I, Kemker R, Part J L, et al. Continual lifelong learning with neural networks: A review[J]. Neural Networks, 2019, 113: 54-71.

[21] Wang L, Zhang X, Su H, et al. A comprehensive survey of continual learning: Theory, method and application[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5362-5383.

[22] Buzzega P, Boschini M, Porrello A, et al. Dark experience for general continual learning: A strong, simple baseline[J]. Advances in Neural Information Processing Systems, 2020, 33: 15920-15930.

[23] Xinwang L. Hyperparameter-free localized simple multiple kernel K-means with global optimum[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7):8566-8576.

[24] Zhou Y, Nezhadarya E, Ba J. Dataset distillation using neural feature regression[J]. Advances in Neural Information Processing Systems, 2022, 35: 9813-9827.

[25] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 4171-4186.

[26] 郝少璞, 刘全, 徐平安, 等. 基于余弦相似度的多模态模仿学习方法[J]. 计算机研究与发展, 2023, 60(6):1358-1372.

[27] Shi H, Xu Z, Wang H, et al. Continual learning of large language models: A comprehensive survey[J]. ACM Computing Surveys, 2024. DOI: 10.1145/3735633.

[28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017. DOI:10.48550/arXiv.1706.03762.

[29] Schneider S, Baevski A, Collobert R, et al. wav2vec: Unsupervised pre-training for speech recognition[J]. arxiv preprint arxiv: 1904.05862, 2019.

[30] 张伟, 钱龙玥, 张林, 等. 多模态持续学习方法研究进展[J]. 数据采集与处理, 2025, 40(5): 1122-1138.

作者简介：

石玥（1999-），女，硕士，研究方向：机器学习、信号处理等，E-mail：1222013922@njupt.edu.cn。

高倩格（2001-），女，硕士研究生，研究方向：机器学习、信号处理等。

季薇（1979-），**通信作者**，女，博士，教授，硕士生导师，研究方向：信号与信息处理、机器学习等，E-mail：jiwei@njupt.edu.cn。

李云（1974-），男，博士，教授，博士生导师，研究方向：机器学习与模式识别等。