基于正样本对比与掩蔽重建的自监督语音表示学习

张文林，刘雪鹏，牛铜，陈琦，屈丹
（信息工程大学信息系统工程学院，河南 郑州 450001）

摘 要：针对现有基于对比预测的自监督语音表示学习方法在训练时需要构建大量负样本，其学习效果依赖于大批次训练，需要耗费大量计算资源的问题，提出了一种仅使用正样本进行语音对比学习的方法，并将其与掩蔽重建任务相结合得到一种多任务自监督语音表示学习方法，在降低训练复杂度的同时提高语音表示学习的性能。其中，正样本对比学习任务，借鉴图像自监督表示学习中 SimSiam 方法的思想，采用孪生网络架构对原始语音信号进行两次数据增强，并使用相同的编码器进行处理，将一个分支经过一个前向网络，另一个分支使用梯度停止策略，调整模型参数以最大化 2 个分支输出的相似度。整个训练过程中不需要构造负样本，可采用小批次进行训练，大幅提高了学习效率。使用 LibriSpeech 语料库进行自监督表示学习，并在多种下游任务中进行微调测试，对比实验表明，所提方法得到的模型在多个任务中均达到或者超过了现有主流语音表示学习模型的性能。

关键词：语音表示；自监督学习；无监督学习；孪生网络
中图分类号：TN912.34
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022142

Self-supervised speech representation learning based on positive sample comparison and masking reconstruction

ZHANG Wenlin, LIU Xuepeng, NIU Tong, CHEN Qi, QU Dan
College of Information System Engineering, Information Engineering University, Zhengzhou 450001, China

Abstract: To solve the problem that existing contrastive prediction based self-supervised speech representation learning methods need to construct a large number of negative samples, and their performance depends on large training batches, requiring a lot of computing resources, a new speech representation learning method based on contrastive learning using only positive samples was proposed. Combined with reconstruction loss, the proposed method could obtain better representation with lower training cost. The proposed method was inspired by the idea of the SimSiam method in image self-supervised representation learning. Using the siamese network architecture, two random augmentations of the input speech signals were processed by the same encoder network, then a feed-forward network was applied on one side, and a stop-gradient operation was applied on the other side. The model was trained to maximize the similarity between two sides. During training processing, negative samples were not required, so small batch size could be used and training efficiency was improved. Experimental results show that the representation model obtained by the new method achieves or exceeds the performance of existing mainstream speech representation learning models in multiple downstream tasks.

Keywords: speech representation, self-supervised learning, unsupervised learning, siamese network

0 引言

语音表示学习是指利用机器学习的方法自动提取语音信号的高层特征表示，该高层特征表示含有音素、说话人、语种等丰富的信息[1]，可用于连续语音识别、说话人识别、语种识别等多种下游任务[2]。自监督语音表示学习利用大量无标注数据学习语音表示，可大大降低下游任务对标注数据的依

收稿日期：2022−04−02；**修回日期：**2022−06−20
基金项目：国家自然科学基金资助项目（No.61673395, No.62171470）
Foundation Item: The National Natural Science Foundation of China (No.61673395, No.62171470)

赖。对于大部分的人类语言，标注数据的获取是十分困难的，一些语言甚至没有书面文字，而无标注语音的获取相对容易得多，因此自监督语音表示学习具有重要的研究意义和应用前景。

目前，常用的自监督语音表示学习方法主要有 3 种[3-4]：基于对比预测的方法、基于掩蔽重建的方法和基于多任务学习的方法。基于对比预测的方法采用区分性模型，以对比预测为辅助任务，通过最小化某种分类损失来得到语音的表示；基于掩蔽重建的方法采用生成式模型，以语音重建为辅助任务，通过最小化某种重建损失来得到语音的表示；基于多任务学习的方法将对比预测任务和语音重建任务相结合，以期达到更好的语音表示效果。从实验结果来看[4]，基于对比预测的方法通常优于基于掩蔽重建的方法，是目前研究语音表示学习最热门的方法。

基于对比预测的方法需要解决的主要问题是模型坍塌问题，即模型所有的表示输出都为一个常数。为了解决这一问题，现有方法往往需要精心设计某种随机采样机制以构造大量的负样本，将其与正样本进行对比，从而避免所有样本的输出都相同。目前，基于对比预测的自监督语音表示学习方法中[5-8]，由于没有语音标注，通常使用相邻的语音帧作为正样本，对来自同一批次的数据进行随机采样以构造负样本。这种正负样本的选择方法没有经过很好的设计，不能保证采样得到的负样本质量，可能会出现错误的正负样本对，导致训练不稳定。具体实现中往往需要通过增大批次大小来保证训练的收敛性和模型的性能，因此，这类方法在训练过程中需要耗费大量的存储和计算资源，计算成本很高。

近年来，众多自监督图像对比表示学习的研究表明[9]负样本并不是训练所必需的，特别是 Chen 等[10]提出的 SimSiam（simple siamese）模型使用简单的孪生网络进行自监督训练，只需要正样本进行对比学习，通过梯度停止技术有效避免了模型坍塌，在 ImageNet 上训练 100 个 epoch 即可达到很好的学习效果。

在语音表示学习领域，目前还没有使用正样本对比学习方法的相关研究。本文将正样本对比学习技术应用于语音表示，并提出了一种结合正样本对比和掩蔽重建的自监督语音表示学习方法。SimSiam 模型[10]采用残差网络（ResNet, residual network）[11]作为表示学习的主干网络；与图像不同，语音信号是一种连续的变长序列，因此本文方法采用 Transformer[12]作为语音表示学习的主干网络。本文所提语音表示学习方法仅需正样本进行对比学习，有效避免了大量负样本带来的模型训练困难问题，结合掩蔽重建任务，在低训练成本下得到了较好的语音表示。本文基于 Librispeech 语料库[13]进行自监督表示学习，并在音素分类、说话人分类等下游任务上进行了测试验证，结果表明与现有模型相比，本文方法得到的模型在所有下游任务中的性能均达到或者超过了现有模型的性能。

1 相关工作

1.1 自监督语音表示学习

目前，自监督语音表示学习的主要方法有基于对比预测的方法、基于掩蔽重建的方法和基于多任务学习的方法 3 种。

基于对比预测的方法主要基于对比预测编码（CPC, contrastive predictive coding）[5]的思想对模型进行自监督学习，利用编码器将语音映射到隐藏的表示空间，要求在表示空间可以对正负样本进行预测分类。这类方法的典型代表是 Facebook 的研究人员提出的 wav2vec 系列模型[6-8]及其相关改进模型[14-15]，其中 wav2vec 2.0 模型[8]已经在连续语音识别等多个下游任务上得到广泛使用，性能达到甚至超过了有监督学习的方法。HuBERT（hidden-unit bidirectional encoder representation from transformers）[14]在 wav2vec 2.0 模型的基础上，通过 K-means 聚类迭代生成离散的伪标签，把预测标签的任务用于模型的训练，该方法可视为对比预测编码的一种改进方法，即通过聚类得到数量固定的负样本进行对比学习。WavLM[15]沿用了 HuBERT 的思想，采用了更灵活的位置编码策略，引入更复杂的加噪和重叠语音数据扩展方法，增大了无标注训练数据的规模，使语音表示模型在 13 个语音相关下游任务中达到最佳性能。上述方法的一个共同缺点是需要采用较大批次进行训练，训练过程中计算复杂度高，容易发生不稳定现象。

基于掩蔽重建的方法主要使用自回归预测编码（APC, autoregressive predictive coding）[16]或 BERT（bidirectional encoder representation from transformer）掩蔽重构任务[17]等对表示模型进行自监督训练。例如，APC 和矢量量化自回归预测编

码（VQAPC, vector-quantized autoregressive predictive coding）[18]采用自回归模型，基于过去语音帧的编码信息来重建未来的语音帧；Mockingjay[19]、TERA（transformer encoder representation from alteration）[20]等模型通过对语音帧进行掩蔽重建来进行模型训练；pMPC（phoneme-based masked predictive coding）[21]在音素级进行掩蔽重构，以期模型学习到的表示能够包含更多的音素类别信息。

基于多任务学习的方法将上述 2 种自监督学习任务进行结合，通过多任务训练的方式得到更有效的语音表示模型。典型代表有 Speech SimCLR[22]，它借鉴了图像自监督表示学习中 SimCLR 模型[23]的对比训练任务，并将其与 TERA 掩蔽重建任务相结合，证明了对比学习任务和掩蔽重建任务相结合的有效性。Zaiem 等[24]探索了多种自监督学习任务相互组合的方法。上述模型在对比学习任务中需要构建大量的负样本对进行对比学习，因此仍需要较大的训练批次，训练复杂度较高。

1.2　基于孪生网络的表示学习

在图像表示学习中，基于孪生网络的表示学习是目前研究的热门方向。孪生网络[25]通过对输入进行 2 种不同增强，以最大化 2 个分支输出的相似性进行训练。然而，如果只是拉近相似样本之间的距离，模型很容易得到一个退化解，即对于所有的样本输出都相同，这就是著名的模型坍塌问题。解决这一问题的一个自然想法是不仅要拉近相似输入之间的距离，也要使不同输入之间的距离变大，换句话说就是不仅要有正样本对，也要有负样本对。这种方法确实解决了模型崩塌的问题，但是也带来了一个新的问题，那就是对负样本对的数量要求较大，只有这样才能训练出足够强的特征表示能力。例如，MoCo（momentum contrast）[26]通过一个记忆库来存储所有样本的特征，从而维护了一个庞大的负样本队列。SimCLR[23]通过 9 000 以上的批次大小来保证有足够数量的负样本供对比学习。

然而，最新的研究结果表明，为了保证图像表示模型不发生坍塌，负样本并不是必需的。例如，BYOL（bootstrap your own latent）[9]基于孪生网络构造了在线网络和目标网络 2 个子网络，2 个子网络分别通过反向传播和动量编码器更新参数，利用在线网络的输出预测目标网络的输出，从而在不使用负样本的情况下避免了模型坍塌。SimSiam[10]进一步去除了动量编码器，将 2 个子网络进行权值共享，对目标网络使用梯度停止策略，从而仅依靠正样本使模型学习到有意义的表示。SimSiam 方法的模型构造和训练过程更简单，只需要小的训练批次即可进行表示学习，本文的自监督语音表示学习中的对比学习任务即受其启发而来。

2　所提模型

2.1　模型结构

基于正样本对比与掩蔽重建的自监督语音表示学习模型如图 1 所示。模型采用孪生网络结构，由 2 个分支组成，其中右分支包含一个用于对输入的语音信号进行编码表示的编码器（Enc）、一个用于根据编码表示对原始语音进行重建的预测器（Pre）；左分支除了编码器和预测器外，还包含一个映射器（Proj）。在模型训练过程中，首先对输入的语音帧 $\boldsymbol{x}$ 进行两路增强得到正样本对 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$，将其分别送入孪生网络的 2 个分支中。左分支对 $\boldsymbol{x}_1$ 进行编码得到特征表示 $\boldsymbol{z}_1$，右分支对 $\boldsymbol{x}_2$ 进行编码得到特征表示 $\boldsymbol{z}_2$。将 $\boldsymbol{z}_1$、$\boldsymbol{z}_2$ 分别送入左、右分支的预测器，得到重建的语音帧 $\boldsymbol{x}_1'$ 和 $\boldsymbol{x}_2'$，分别计算重建损失函数。对于语音表示 $\boldsymbol{z}_1$，还将其送入左分支的映射器得到输出 $\boldsymbol{p}_1=\mathrm{Proj}(\boldsymbol{z}_1)$ 来对右分支的语音表示 $\boldsymbol{z}_2$ 进行预测，映射器的输出 $\boldsymbol{p}_1$ 的维度与语音表示 $\boldsymbol{z}_1$、$\boldsymbol{z}_2$ 相同，计算 $\boldsymbol{p}_1$ 与 $\boldsymbol{z}_2$ 之间的余弦距离作为对比预测损失函数，并对其进行最小化以使两路输出更相似。将 $\boldsymbol{z}_1$、$\boldsymbol{z}_2$ 的重建损失函数及对比预测损失函数进行求和作为最终的损失函数，采用梯度下降法对其进行优化，从而实现语音表示的多任务学习。

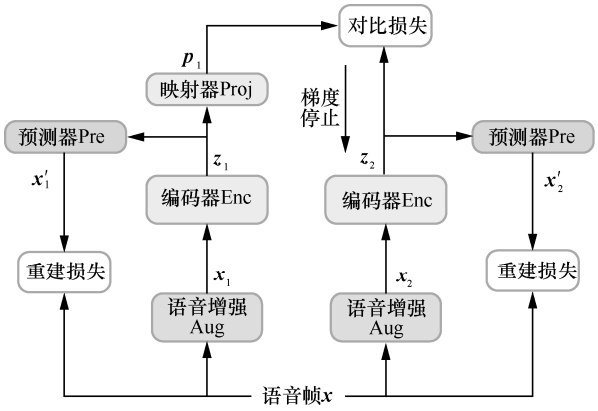


图 1　基于正样本对比与掩蔽重建的自监督语音表示学习模型

在图像的正样本表示学习中，采用 ResNet[11] 作为编码器。与图像不同，语音信号是一个连续的变长序列，具有长时相关性，因此在本文方法中左右分支的编码器均采用多层 Transformer 编码器[12] 的堆叠实现。其中每一个编码器层由一个多头自注意力层和一个前向层组成，将编码器的最后一层输出视为最终的语音表示。模型中的预测器和映射器均采用两层前向网络，仅在语音表示学习（即预训练）阶段使用，表示学习结束后仅保留编码器，将得到的语音表示作为下游任务的输入特征。

2.2 数据增强

针对语音数据，本文采用加入高斯噪声、时间掩蔽、频率掩蔽 3 种方式对输入的语音进行随机扰动获得增强后的语音数据，以得到孪生网络 2 个分支的输入数据。考虑到在测试阶段并不会对输入模型的语音数据进行增强，为了减少训练与测试的不一致，对左右 2 个分支均以概率 p_a 进行随机增强。

具体地，对于一条输入语音帧 $\boldsymbol{x}$，根据[0,1]上的均匀分布进行采样得到 $p \in [0,1]$，当 $p \geqslant p_a$ 时进行语音增强，否则不进行语音增强。增强后的语音数据为

$$\boldsymbol{x}' = \begin{cases} \boldsymbol{x} & ,\ p < p_a \\ \mathrm{Aug}(\boldsymbol{x}) & ,\ p \geqslant p_a \end{cases} \tag{1}$$

其中，$\mathrm{Aug}(\boldsymbol{x})$ 表示对输入语音 $\boldsymbol{x}$ 进行随机扰动，实验中通过 Specaugment 工具箱[27]实现。

2.3 正样本对比损失函数计算及模型训练方法

在模型训练过程中，左右 2 个分支的编码器和预测器共享相同的权值。最终总的损失函数由重建损失和预测损失两部分组成。其中，重建损失为使用左右 2 个分支编码输出 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_2$ 对原始语音帧 $\boldsymbol{x}$ 进行重建的误差。借鉴 TERA 模型[20]中的做法，本文使用 L1 距离来衡量语音帧的重建误差，总重建损失为

$$L_{\mathrm{recon}} = l_1\left(\boldsymbol{x}, \boldsymbol{x}_1'\right) + l_1\left(\boldsymbol{x}, \boldsymbol{x}_2'\right) \tag{2}$$

其中，l_1 为 L1 距离，$\boldsymbol{x}_1'$ 和 $\boldsymbol{x}_2'$ 分别为左右 2 个分支的重建语音帧。

预测损失由左右 2 个分支的语音特征表示 $\boldsymbol{z}_1$、$\boldsymbol{z}_2$ 进行相互预测计算得到。由于 $\boldsymbol{z}_1$、$\boldsymbol{z}_2$ 分别对应同一个输入语音帧 $\boldsymbol{x}$ 的不同增强 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$，可视为正样本对，因此在训练预测任务时只需要最大化预测值与预测目标之间的余弦相似度即可。

为防止模型坍塌，在反向传播过程中，对右分支执行梯度停止操作，不对其模型参数进行梯度计算。图像自监督学习方法 SimSiam[10]从实验验证和理论证明 2 个角度证明了梯度停止在仅采用正样本进行对比学习中的重要性。从理论上看，梯度停止操作相当于将语音表示视为隐藏变量，引入除模型参数之外的第二组参数，采用 EM（expactation maximization）算法，通过两组参数的迭代更新得到表示模型参数的近似最优估计。

综上所述，采用梯度停止技术后，根据 $\boldsymbol{z}_1$ 对 $\boldsymbol{z}_2$ 进行预测的损失函数为

$$l_{\mathrm{pred}}(\boldsymbol{z}_1, \boldsymbol{z}_2) = -\cos\left(\mathrm{Proj}(\boldsymbol{z}_1), \mathrm{stopgrad}(\boldsymbol{z}_2)\right) \tag{3}$$

其中，$\mathrm{stopgrad}(\cdot)$ 表示梯度停止操作；$\mathrm{Proj}(\boldsymbol{z}_1)$ 表示 $\boldsymbol{z}_1$ 经过预测器后的输出；$\cos(\cdot)$ 表示矢量间的余弦距离，其计算式为

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}}{\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2} \tag{4}$$

其中，$\|\cdot\|_2$ 表示 L2 范数。

实际中由于左右 2 个分支的编码器参数完全相同，为了更高效地利用训练样本，可同时计算根据 $\boldsymbol{z}_2$ 对 $\boldsymbol{z}_1$ 进行预测的损失函数，得到对称的总预测损失函数为

$$L_{\mathrm{pred}} = \frac{1}{2}\left[l_{\mathrm{pred}}(\boldsymbol{z}_1, \boldsymbol{z}_2) + l_{\mathrm{pred}}(\boldsymbol{z}_2, \boldsymbol{z}_1)\right] \tag{5}$$

结合式(2)和式(5)，最终训练总损失函数为

$$L = L_{\mathrm{recon}} + L_{\mathrm{pred}} \tag{6}$$

TERA 模型[20]中，将重建损失函数和对比预测损失函数的权重均设置为 1。采用上述损失函数进行梯度下降训练，最终将学习到的编码器 Enc 作为语音表示模型。

3 实验

3.1 实验设置

实验中使用公开数据集 Librispeech[13]训练语音表示模型。为了验证不同数据量下模型的性能，分别使用其中的 100 h 无标注数据（train-clean-100）和 960 h 无标注数据（包括 train-clean-100、train-clean-360、train-other-500）进行自监督训练。编码器为 3 层 Transformer 编码器，隐含层维度为 768，注意力头为 12 个，前向层维度为 3 072，dropout 概率为 0.1。预测头和投影头均由 2 个线性层构成，隐含层维度为 768。编码器的输入为 80 维美尔谱，帧长为 25 ms，帧移为 10 ms。

将最后一层编码器的输出作为最终的语音表示，后接一个任务相关的分类器或预测器用于下游任务。在下游任务训练中，语音表示模型有 2 种使用方式：一种是将编码器参数冻结，仅作为特征提取器使用；另一种是将编码器参数与下游任务相关的模型参数共同在下游任务的标注数据上进行微调训练。预训练模型经过微调后更适应具体下游任务的需要，因此在测试中具有更好的效果。

实验中测试了音素分类、说话人分类及连续语音识别等下游任务的性能。其中，音素分类任务遵循 CPC[5]的实验设置，使用 CPC 给出的数据集划分方式将 train-clean-100 划分为训练集和测试集，分别比较了使用单层线性层的分类器和含一个隐含层的分类器的识别性能，下文中将上述 2 种分类器的实验结果分别记为 lib-plin 和 lib-phid；同时实验中也测试了在 TIMIT（Texas Instruments/Massachusetts Institute of Technology）数据集[28]上使用单层线性层作为分类器的实验结果，记为 timit-plin。训练过程中预训练模型参数被冻结。

在说话人分类任务中，同样遵循 CPC[5]的实验设置，使用 CPC 给出的数据集划分方式将 train-clean-100 划分为训练集和测试集，使用单层线性网络作为分类器，测试模型的帧级说话人分类准确率，记为 spk-fr。训练过程中预训练模型参数被冻结。

在连续语音识别任务中，将语音表示作为特征输入，以 Conformer 模型[29]为声学模型，以 Transformer 解码器为语言模型构建连续语音识别系统，分别使用 Librispeech 的 100 h 和 960 h 的有标注数据作为训练集对模型进行训练，测试系统在 test-clean 子集上的词错误率（WER, word error rate）。训练过程中预训练模型参数与语音识别系统一起微调训练。

3.2　100 h 无标注数据实验结果

本节使用 train-clean-100 对表示模型进行预训练，测试其在音素分类和说话人分类任务上的性能。同时采用加入高斯噪声、时间掩蔽和频率掩蔽 3 种数据增强方式，对图 1 中左右 2 个分支的输入均以概率 p_a 进行随机增强。模型训练过程中使用 4 个 1080Ti GPU，总的训练批次大小为 8（约 12 s 语音数据），训练总步数为 200 000 步。在训练下游任务时，语音表示模型的参数被冻结，仅作为特征提取器使用。

首先，对不同的数据增强概率 p_a 进行了对比实验，分别测试了 p_a=0.5、p_a=0.8、p_a=1.0 时的模型性能。不同数据增强概率时模型在下游任务中的实验结果如表 1 所示。表 1 中还给出了仅对左分支输入进行增强（单路增强）的效果。

表 1　不同数据增强概率时模型在下游任务中的实验结果

增强概率	timit-plin	lib-plin	lib-phid	spk-fr
单路增强	70.86%	70.97%	79.27%	99.80%
p_a=1 .0	66.97%	64.81%	73.50%	97.59%
p_a=0.5	71.39%	71.25%	79.77%	99.76%
p_a=0.8	70.27%	69.85%	78.50%	99.62%

从表 1 可知，在音素分类任务上，$p_a = 0.5$ 时模型性能最佳；在说话人分类任务中，单路增强的模型结果最佳，$p_a = 0.5$ 时模型准确率与之接近，仅相差 0.04。因此，在后面的实验中，取 $p_a = 0.5$ 进行增强。

进一步地，分别测试了本文模型在仅使用重建损失、仅使用正样本对比损失及使用多任务学习时的性能，同时还对比测试了使用相同训练数据的 Mockingjay、TERA 模型的性能。100 h 无标注数据下的实验结果如表 2 所示。

表 2　　100 h 无标注数据下的实验结果

测试模型	timit-plin	lib-plin	lib-phid	spk-fr
CPC	64.8%	64.6%	72.5%	97.4%
Mockingjay[19]	64.7%	60.1%	75.3%	83.4%
TERA[20]	65.6%	65.2%	77.3%	98.9%
pMPC[21]	68.1%	67.3%	78.8%	99.5%
sia100-rec	69.66%	70.52%	78.84%	99.65%
sia100-con	49.72%	46.32%	54.02%	75.95%
sia100-mt	71.39%	71.25%	79.77%	99.76%

表 2 中，sia 表示本文模型，100 表示训练数据为 100 h，rec、con 和 mt 分别表示仅使用重建损失、仅使用正样本对比损失和使用多任务学习时的实验结果。例如，sia100-rec 表示本文模型在 100 h 训练数据、仅使用重建损失时的实验结果。

仅使用重建损失的模型（sia100-rec）可以看作对 TERA 模型进行了数据增强，从表 2 可知，其在各下游任务中的性能均优于 TERA 模型。仅使用正样本对比损失的模型（sia100-con）能够学习到一定的信息，但是性能相对较差，本文推测这是由于模型在训练过程中仅对单个语音帧进行对比，可供利用的信息较少，导致模型无法得到充分训练。使用正样本对比和重建损失进行多任务学习的模型（sia100-mt）在所有下游任务的测试中性能均达到最佳，验证了本文模型的有效性。

3.3 960 h 训练数据实验结果

本节进一步在 960 h 无标注数据上对模型进行训练，参数设置与 3.2 节实验基本相同。总的训练批次大小设置为 8（约 12 s 语音数据），训练总步数为 1000 000。实验结果如表 3 所示。

从表 3 可知，本文模型（sia960-mt）在所有下游任务中性能均优于 TERA 模型。在说话人分类任务（spk-fr）和 TIMIT 数据集上的音素分类任务（timit-plin）中，本文模型（sia960-mt）的准确率超过了 wav2vec 2.0-Base 模型；在 librispeech 的音素分类任务（lib-plin 和 lib-phid）中，性能略低于 wav2vec 2.0-Base 模型。出现这一结果的原因在于，本文模型参数量较少，学习的语音表示复杂性相对大模型来说较低，具有更好的线性可分性，因此音素分类任务在单层线性分类器上表现更优，而在使用多层线性层作为分类器时，测试的准确率相比 wav2vec 2.0-Base 模型略低。

本文进一步增大了模型训练的批次大小，将总的训练批次增大到 32，训练总步数为 1 000 000 步，表示为 sia960-mt-bs32。从实验结果看，增大训练批次能进一步提高语音表示模型的性能，在所有下游任务测试中均取得了更好的结果。

由于本文模型参数数量仅为 wav2vec 2.0-Base 模型的 $\frac{1}{4}$，且仅使用正样本进行对比学习，因此可以在仅使用 4 个 2080Ti GPU、总的训练批次大小为 32（相当于 48 s 语音数据）时即取得较好的训练效果，模型训练所需时间约为 9.5 天。与之相比，相同训练数据量条件下，wav2vec 2.0-Base 模型在训练过程中使用了 64 个 V100 GPU，总的训练批次大小为 1.6 h。如表 4 所示，在不考虑不同 GPU 性能的情况下，本文模型的预训练时间相当于使用单个 GPU 预训练 38 天，wav2vec 2.0-Base 模型则相当于使用单个 GPU 预训练 102 天。因此，本文方法训练更高效，所需计算资源更少，实际中更有利于模型的推广应用。表 4 中，下游任务训练速度和推理速度为归一化结果。在推理时间上，使用单层线性层进行音素分类任务的测试时，wav2vec 2.0-Base 模型在测试集上的推理时间为 3.8 min，本文模型仅需要 1.7 min，推理速度是 wav2vec 2.0-Base 模型的 2.2 倍。

3.4 语音识别

连续语音识别任务使用 ESPnet 工具箱[30]实现，使用 960 h 无标注数据对表示模型进行预训练，本节分别在 100 h 和 960 h 有标注数据条件下对连续语音识别模型进行微调，以分别模拟低资源和高资源训练条件，分别测试模型在 test-clean 子集上的 WER，测试结果如表 5 所示。声学模型采用 Conformer 模型，Trans 表示语言模型采用 Transformer 解码器结构，4-gram 表示采用四元文法语言模型。

从表 5 的实验结果可知，在使用 100 h 有标注数据进行微调时，与相同规模参数量的模型相比，sia960-mt 与 Speech SimCLR 的 WER 最低，为 5.7%；进一步将训练批次大小增加至 32 后，sia960-mt-bs32 的 WER 进一步降低，为 5.5%。这一实验结果与相同训练数据量（100 h 有标注数据）下的 wav2vec 2.0-Base 和 HuBERT-Base 相比还有一定的差距。使用 960 h 有标注数据进行微调后，sia960-mt 与 wav2vec 2.0-Base 和 HuBERT-Base 模型相比，WER 仅差 0.3%和 0.2%，而参数量仅约为它们的 $\frac{1}{4}$。增大训练批次大小后，sia960-mt-bs32 模型的性能进一步提高，WER 降低到 2.3%。

表 3 **960 h 无标注数据的测试结果**

测试模型	参数量/个	timit-plin	lib-plin	lib-phid	spk-fr
Mockingjay[19]	22×10^6	58.4%	67.0%	79.1%	99.3%
TERA[20]	222×10^6	70.0%	71.2%	80.2%	99.2%
wav2vec2.0-Base[8]	952×10^6	73.26%	75.89%	85.54%	99.40%
sia960-mt	232×10^6	74.34%	74.80%	82.40%	99.69%
sia960-mt-bs32	232×10^6	75.33%	76.36%	83.17%	99.78%

表 4 **本文模型与 wav2vec 2.0-Base 模型参数量、训练资源比较**

模型	参数量/个	训练资源	预训练时间/天（单个 GPU）	下游任务训练速度	推理速度
本文模型	23×10^6	2080Ti GPU×4	38	1	1
wav2vec2.0-Base[8]	95×10^6	V100 GPU×64	102	4.3	2.2

表 5　语音识别任务测试结果

模型	参数量/个	语言模型	标注数据/h	WER
wav2vec-Large[6]	33×10^6	Trans	100	6.9%
TERA[20]	22×10^6	Trans	100	6.0%
Speech SimCLR[22]	30×10^6	Trans	100	5.7%
sia960-mt	23×10^6	Trans	100	5.7%
sia960-mt-bs32	23×10^6	Trans	100	5.5%
wav2vec 2.0-Base[8]	95×10^6	4-gram	100	3.4%
HuBERT-Base[14]	95×10^6	4-gram	100	3.4%
wav2vec 2.0-Base[8]	95×10^6	Trans	960	2.1%
HuBERT-Base[14]	95×10^6	Trans	960	2.2%
sia960-mt	23×10^6	Trans	960	2.4%
sia960-mt-bs32	23×10^6	Trans	960	2.3%

3.5　SUPERB 基准测试

SUPERB[4]是一个语音处理通用性能基准排行榜，旨在为语音社区提供一个标准和全面的测试平台，用于评估预训练模型在涵盖语音所有方面的各种任务上的可推广性，通过 10 多个任务来测试语音表示对语音内容、说话人、语义和副语言 4 个方面的表达能力。

本节主要在 SUPERB 的 6 个任务上对本文模型进行测试，具体如下。

音素识别（PR, phoneme recognition）任务。使用 LibriSpeech 数据集的 train-clean-100、dev-clean、test-clean 子集分别作为训练集、验证集、测试集，测试的评价指标为音素错误率（PER, phone error rate）。

关键词识别（KS, keyword spotting）任务。使用 Speech Commands v1.0 数据集，包含 10 种关键词、静音和未知，评价指标为准确率（ACC, accuracy）。

意图分类（IC, Intent classification）任务。在话语级对说话人的意图进行分类，使用 Fluent Speech Commands 数据集[31]，评价指标为 ACC。

说话人辨认（SID, speaker identification）任务。在话语级进行分类，使用 VoxCeleb1 数据集[32]进行训练和测试，评价指标为 ACC。

实例查询口语术语检测（QbE, query by example spoken term detection）任务。采用 QUESST 2014 [33]挑战赛中的英语子集，评估指标是最大项加权值（MTWV, maximum term weighted value），能更好地平衡漏检和虚警。

说话人确认（SV, speaker verification）任务。使用 VoxCeleb1 数据集，采用 x-vecter[34]作为下游模型，评价指标为等错误率（EER, equal error rate）。

实验中采用 SUPERB 默认的测试方法对上述 6 个任务进行测试，其中预训练模型仅作为前端的特征提取器，参数保持冻结。

表 6 中展示了多种自监督预训练模型在 SUPERB 基准上的测试结果。从表 6 中可以看出，当使用本文自监督语音表示模型学习的表示作为下游任务的输入时，在绝大多数任务上相比 log-Mel 谱均有明显提升。本文模型与参数量相近的模型相比，在所有任务上性能都达到最佳；与参数量大得多的 wav2vec 2.0-Base 和 HuBERT-Base 模型相比，在多数任务上也有相当的表现，特别是在 SID 任务上，性能甚至超过了 wav2vec 2.0-Base 和 HuBERT-Base 模型。

4　结束语

本文提出了一种新的自监督语音表示学习方法，该方法基于孪生网络结构，将仅使用正样本的对比学习任务和掩蔽重建任务相结合，以多任务学习的方式实现自监督语言表示学习。与现有方法相比，本文方法仅使用了正样本进行对比学习，通过

表 6　多种自监督预训练模型在 SUPERB 基准上的测试结果

模型	可训练参数量/个	PR-PER↓	KS-ACC↑	IC-ACC↑	SID-ACC↑	QbE-MTWV↑	SV-EER↓
log-Mel 谱	0	82.01%	8.63%	9.10%	8.5×10^{-6}	0.005 8	9.56%
Mockingjay[19]	19×10^6	58.88%	82.37%	30.29%	35.84%	7.0×10^{-4}	10.91%
TERA[20]	22×10^6	47.53%	88.09%	48.8%	58.67%	8.7×10^{-5}	16.49%
wav2vec 2.0-Base[8]	95×10^6	28.37%	92.31%	58.34%	45.62%	8.8×10^{-4}	9.69%
HuBERT-Base[14]	95×10^6	6.85%	95.98%	95.94%	64.84%	0.075 9	7.22%
sia960-mt	23×10^6	46.58%	89.81%	66.54%	65.13%	0.017 8	9.23%
sia960-mt-bs32	23×10^6	44.40%	91.37%	66.70%	68.59%	0.018 1	9.52%

梯度停止策略防止模型坍塌，仅需要较小的批次大小即可进行训练。在音素分类、说话人分类和连续语音识别等下游任务中的实验测试表明，与 TERA 等同等参数量下的其他模型相比，本文方法具有更好的性能；同时，与 wav2vec2.0-Base 等参数量大得多的表示学习方法相比，本文方法在大大降低模型训练的存储和计算开销的同时达到甚至超过了其性能，具有良好的应用推广价值。

参考文献：

[1] 陈虹洁. 面向低资源场景的语音表示学习及其应用[D]. 西安: 西北工业大学, 2018.
CHEN H J. Low-resource speech representation learning and its applications[D]. Xi'an: Northwestern Polytechnical University, 2018.

[2] 朱毅. 基于深度学习的表示学习算法研究[D]. 合肥: 合肥工业大学, 2018.
ZHU Y. Research on deep learning-based representation learning algorithms[D]. Hefei: Hefei University of Technology, 2018.

[3] 刘雪鹏, 张文林. 自监督语音表示学习综述[C]//2021 年第十六届全国人机语音通信学术会议录, 北京: 中国中文信息学会, 2021: 284-293.
LIU X P, ZHANG W L. An overview of self-supervised speech representation learning[C]//2021 National Conference on Man-Machine Speech Communication 2021. Beijing: Chinese Information Processing Society of China, 2021: 284-293.

[4] YANG S W, CHI P H, CHUANG Y S, et al. SUPERB: speech processing universal performance benchmark[C]//Proceedings of Interspeech 2021. Piscataway: IEEE Press, 2021: 1194-1198.

[5] OORD A V D, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv Preprint, arXiv: 1807.03748, 2018.

[6] SCHNEIDER S, BAEVSKI A, COLLOBERT R, et al. wav2vec: unsupervised pre-training for speech recognition[C]//Proceedings of Interspeech 2019. Piscataway: IEEE Press, 2019: 3465-3469.

[7] BAEVSKI A, SCHNEIDER S, AULI M. vq-wav2vec: self-supervised learning of discrete speech representations[J]. arXiv Preprint, arXiv: 1910.05453, 2019.

[8] BAEVSKI A, ZHOU H, MOHAMED A, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations[J]. Advances in Neural Information Processing Systems, 2020, 33: 12449-12460.

[9] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 21271-21284.

[10] CHEN X L, HE K M. Exploring simple Siamese representation learning[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 15745-15753.

[11] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2017: 6000-6010.

[13] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2015: 5206-5210.

[14] HSU W N, TSAI Y H H, BOLTE B, et al. HuBERT: how much can a bad teacher benefit ASR pre-training? [C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 6533-6537.

[15] CHEN S Y, WANG C Y, CHEN Z Y, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing[J]. arXiv Preprint, arXiv: 2110.13900, 2021.

[16] CHUNG Y A, HSU W N, TANG H, et al. An unsupervised autoregressive model for speech representation learning[C]//Proceedings of Interspeech 2019. Piscataway: IEEE Press, 2019: 146-150.

[17] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.[S.L.:s.n], 2019: 4171-4186.

[18] CHUNG Y A, TANG H, GLASS J. Vector-quantized autoregressive predictive coding[C]//Proceedings of Interspeech 2020. Piscataway: IEEE Press, 2020: 3760-3764.

[19] LIU A T, YANG S W, CHI P H, et al. Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders[C]//Proceedings 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 6419-6423.

[20] LIU A T, LI S W, LEE H Y. TERA: self-supervised learning of transformer encoder representation for speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2351-2366.

[21] YUE X H, LI H Z. Phonetically motivated self-supervised speech representation learning[C]//Proceedings of Interspeech 2021. Piscataway: IEEE Press, 2021: 746-750.

[22] JIANG D W, LI W B, CAO M, et al. Speech SimCLR: combining contrastive and reconstruction objective for self-supervised speech representation learning[C]//Proceedings of Interspeech 2021. Piscataway: IEEE Press, 2021: 1544-1548.

[23] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//2021 International conference on machine learning. New York: PMLR, 2020: 1597-1607.

[24] ZAIEM S, PARCOLLET T, ESSID S. Pretext Tasks selection for multitask self-supervised speech representation learning[J]. arXiv Preprint, arXiv: 2107.00594, 2021.

[25] CHICCO D. Siamese neural networks: an overview[J]. Artificial Neural Networks, 2021: 73-94.

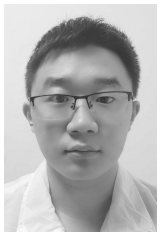
[26] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9726-9735.

[27] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: a simple data augmentation method for automatic speech recognition[C]//Proceedings

of Interspeech 2019. Piscataway: IEEE Press, 2019: 2613-2617.
[28] GAROFOLO J S, LAMEL L F, FISHER W M, et al. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1[R]. NASA STI/Recon Technical Report N, 1993.
[29] GULATI A, QIN J, CHIU C C, et al. Conformer: convolution-augmented transformer for speech recognition[C]//Proceedings of Interspeech 2020. Piscataway: IEEE Press, 2020: 5036-5040.
[30] WATANABE S, HORI T, KARITA S, et al. ESPnet: end-to-end speech processing toolkit[C]//Proceedings of Interspeech 2018. Piscataway: IEEE Press, 2018: 2207-2211.
[31] LUGOSCH L, RAVANELLI M, IGNOTO P, et al. Speech model pre-training for end-to-end spoken language understanding[C]//Proceedings of Interspeech. Piscataway: IEEE Press, 2019: 814-818.
[32] NAGRANI A, CHUNG J S, XIE W D, et al. Voxceleb: large-scale speaker verification in the wild[J]. Computer Speech & Language, 2020, 60: 101027.
[33] ANGUERA X, RODRIGUEZ-FUENTES L J, BUZO A, et al. QUESST2014: evaluating query-by-example speech search in a zero-resource setting with real-life queries[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2015: 5833-5837.
[34] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: robust DNN embeddings for speaker recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 5329-5333.

[作者简介]

张文林（1982− ），男，湖北黄冈人，博士，信息工程大学副教授，主要研究方向为语音信号处理、语音识别、机器学习。



刘雪鹏（1996− ），男，山东泰安人，信息工程大学硕士生，主要研究方向为智能信息处理、无监督学习、语音表示学习。

牛铜（1984− ），男，河南安阳人，博士，信息工程大学副教授，主要研究方向为深度学习、语音信号处理和语音识别。

陈琦（1974− ），男，河南郑州人，信息工程大学副教授，主要研究方向为语音信号处理、语音识别和音频水印。

屈丹（1974− ），女，吉林九台人，博士，信息工程大学教授，主要研究方向为机器学习、深度学习和语音识别。