arXiv:2304.03940v1 [cs.LG] 8 Apr 2023

无监督使用矢量量化的语音表征池化

*Jeongkyun Park**、*Kwanghee Choi**、*Hyunjun Heo**、*Hyung-Min Park*

西江大学{park32323,juice500,hteconn,hpark}
@sogang.ac.kr

摘要

随着大型自监督模型所产生的通用语音表征不断涌现，将单一模型应用于多个下游任务已成为一种事实上的常见做法。然而，池化问题依然存在：语音表征的长度本质上具有可变性。尽管简单的平均池化忽略了语音的特性（例如音素长度各不相同），它仍常被采用。因此，我们设计了一种新颖的池化方法，通过矢量量化将声学上相似的表征压缩在一起；与基于注意力的池化不同，该方法无需额外训练。此外，我们在多种自监督模型上评估了各种无监督池化方法。我们汇集了分散于语音和文本领域的多种方法，并在多项任务上进行评估：关键词检测、说话人识别、意图分类和情感识别。最后，我们通过定量和定性分析，将所提出的方法与有监督池化方法进行比较。

索引词：池化、矢量量化、语音表征、自监督模型、wav2vec 2.0

1. 引言

wav2vec 2.0 [1] 凭借基于 Transformer 的自监督模型所展现出的惊人效果，在语音处理领域掀起热潮 [2]。自监督学习利用大量无标签数据，而无标签数据比有标签数据容易获取得多。任务设计有多种形式，其中最常见的一种是对输入进行掩码，并训练神经网络根据周围信息重建原始输入 [1, 3]。这种任务定义使神经网络能够建模输入分布，从而生成可用于各种下游任务的通用表征。

构建通用的自监督模型后，接下来会进行有监督训练，根据具体需求微调模型。通常会将倒数第二个全连接层接到冻结的共享模型上，作为预测头 [2]。然而，语音和文本的输入长度往往各不相同，这是序列数据的基本特征。随着输入规模增大，Transformer 的输出表征长度也会线性增加，而最终的预测头需要固定大小的输入表征。这就引出了表征池化问题：面对数量可变的序列表征，需要对其进行汇总，使其具有固定大小。

在自然语言处理文献中，池化任务通常被视为无监督或释义句子嵌入问题 [4, 5], ，尤其是在词嵌入在各种任务中取得成功之后。其目标是生成通用的句子嵌入，不受各个下游任务的影响。从简单地对词表征取平均开始 [4], ，人们提出了许多非参数化方法，例如在取平均时考虑词频 [5], 、对白化表征 [6, 7], ，或利用奇异值变换 [8]。然而，现有方法通常依赖底层分词器，因此难以直接应用于语音。此外，简单地对语音表征取平均存在一个根本问题：每个音段的长度不同，因此元音相较于辅音会被过度加权。

与此同时，人们提出了许多池化方法来汇总逐帧语音表示，尤其用于说话人识别。例如，统计池化（SP） [9, 10] 会将表征的一阶和二阶统计量拼接起来。此外，注意力层常用于确定哪些表征对各项下游任务更为重要 [11–13]。然而，针对无监督池化方法的研究仍然有限。除 SP 外，现代方法通常需要大量参数，并依赖标注样本。相比之下，无监督池化完全省去了训练过程，更贴近通用语音嵌入的理念。

为避免参数化，同时使方法更适用于语音任务，我们提出使用向量量化（VQ）替代分词器。数十年来，研究人员一直在探索 VQ [14],，它也逐渐应用于现代自监督模型，例如 vq-wav2vec [15, 16]。VQ 的核心是将实值向量转换为可计数的索引，并通过精心选择的质心对潜在嵌入空间进行划分。鉴于频率、振幅和共振峰等语音的关键组成要素都嵌入在模型表征中 [17], ，我们预期这些簇能够成功聚集相似的音段表征，使它们具有相同的簇索引。我们设计了多种直观的方法，利用 VQ 有效汇总语音表征，并通过实验展示了这些方法在广泛场景下的有效性。

为进一步推动无监督池化方法的研究，我们设计了一套基准测试，用于评估多种骨干网络上的各类池化方法，并覆盖广泛的任务，包括关键词检测 [18], 说话人识别 [19], 意图分类 [20], 和情感识别 [21]。为完全避免使用监督信号，我们紧密遵循近期对比学习文献中的做法 [22, 23] ，通过最近邻方法直接评估池化表示的有效性 [24]。此外，我们对现有的无监督句子嵌入方法进行修改，以比较它们在下游任务中的表现。我们还通过将本方法与基于监督学习的参数化方法进行比较、直接可视化权重，以及探索基准测试评估的不同设置，开展多项分析，以准确展现本方法的有效性和行为特征。

* 贡献相同。

2. 方法

本节简要概述 vq-wav2vec [16] 和 wav2vec 2.0 [1,25] 中使用的矢量量化方法，详见第2.1节。随后，我们讨论利用量化方法的多种方式。第2.2节介绍合并相似表示的思路，而第2.3节则利用量化结果估计每种表示的先验概率。

2.1. 预备知识：矢量量化

在介绍我们的方法之前，我们先简要介绍 wav2vec 2.0 [1] 的内部结构及其矢量量化器模块。将原始语音输入记为 $\mathbf{x}$，将一维卷积特征得到的潜在语音表征记为 $\mathbf{Z}$，将 Transformer 得到的上下文表征记为 $\mathbf{C}$，并将量化表示记为$\mathbf{Q}$:

$$\mathbf{Z} = \texttt{CNN}(\mathbf{x}) \in \mathbb{R}^{T \times F}, \tag{1}$$

$$\mathbf{C} = \texttt{transformer}(\mathbf{Z}) \in \mathbb{R}^{T \times F}. \tag{2}$$

$$\mathbf{Q} = \texttt{quantizer}(\mathbf{Z}) = \{q_t^g\} \in [V]^{T \times G}, \tag{3}$$

其中， G 是码本组的数量；共享码本通常能以更高的效率取得相近的性能

[16] i $1 \le i \le$ 。我 V 们将介于两者之间的整数记为 $[V]$。^{t} 第 - 帧的每个表征由 G 个整数组成，其中每个整数为 $1 \le q_t^g \le V$， V 是质心的数量。因此，对于逐帧语音表征 $\mathbf{Z}_t \in \mathbb{R}^F$，其中 t 是帧索引，量化器将 $\mathbf{Z}_t$ 映射为 G 维整数（$q_t^1, q_t^2, \cdots, q_t^G$）。量化结果旨在将彼此接近的表征映射为相同的索引，因此当声学差异可以忽略时，quantizer(**Z**) $=$ quantizer(**Z** $+ \epsilon$) ϵ。

2.2. 通过量化结果进行 Squash

为便于记号表示，我们将池化视为帧表示的加权和$\mathbf{C}_t$，其中标量 w_t:

$$\bar{\mathbf{C}} = \textstyle\sum_{t=1}^{t=T} \zeta w_t \mathbf{C}_t. \tag{4}$$

其中，归一化常数 ζ 用于确保权重之和等于 1，即，$\sum_{t=1}^{t=T} \zeta w_t = 1$。

例如，对于平均池化（AP），

$w_t = 1$ ζ =和 。我们 $1/T$ 观察到，AP 将每一帧视为同等重要。然而，语音信号的长度与其重要性并不成正比。例如，即使存在一段很长的静音，AP 也会将其纳入平均计算，因此得到的表征会严重偏向静音表征。此外，元音和辅音的固有时长不同，因此表征会更多地拟合元音。

因此，我们提出一种基于 VQ 的新型池化方法，将相似的帧聚类到各个分区中，其思路借鉴了连接主义时序分类（CTC）声学模型的贪心解码 [26]。通过对.....进行平均对于分区，该方法实际上会对较大分区内的表征施加惩罚:

$$\bar{\mathbf{C}}_{\text{Squash}} = \frac{1}{|\mathcal{S}(\mathbf{Q})|} \sum_{P \in \mathcal{S}(\mathbf{Q})} \frac{1}{|P|} \sum_{t \in P} \mathbf{C}_t. \tag{5}$$

$\mathcal{S}$： $[V]^{T \times G} \rightarrow 2^{[V]^{T \times G}}$ 是压缩函数，其中我们将集合 A 的超集记作 2^A。压缩函数返回帧索引的分区，每个分区都通过压缩相似的表征构成。内层分数用于对单个分区内的相似表征进行平均（

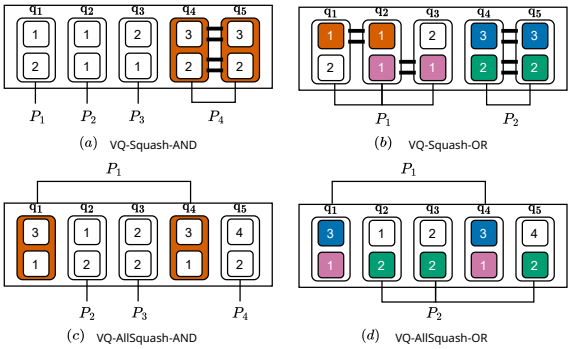

图 1：量化表征的分区方法。

其中包含相似的表征（$1/|P|$），外层分数用于跨分区进行平均。我们从两个角度探讨压缩操作，相关总结见图 1。

首先，可以通过两种方式定义不同步骤中的两个表征$\mathbf{C}_i$和$\mathbf{C}_j$是否相等。我们可以检查$\mathbf{C}_i$和$\mathbf{C}_j$的所有索引（AND），也可以允许部分匹配（或）。与 AND 相比，或旨在利用每个量化组合并更多相邻样本，同时容忍一定误差。

此外，还需要定义应合并哪些时间帧。我们提出一种称为 Squash 的策略，它会对长序列中反复出现的特征进行惩罚，这与 CTC 类似，后者会将相邻的相同值压缩。另一方面，我们发现，平均池化以无序方式进行。因此，可以将 Squash 扩展为对所有出现次数取平均，从而惩罚出现频率高于其他特征的特征。我们将这种策略称为 AllSquash。

2.3. 基于量化结果的先验概率估计

我们扩展 AllSquash 的思路，以估计每个量化表示出现的先验概率 $P(q)$。对于句子嵌入，通常使用平滑逆频率（SIF） [5] 来降低高频词的权重。给定一个词 W 和一个超参数 a，权重为 $w_i = a/(a + p(W))$。 [5] 指出，SIF 与词频-逆文档频率（TF-IDF） [27] 以及 word2vec [28]的子采样概率直接相关。

我们通过使用向量量化（VQ）将 SIF 扩展到语音表征。为此，我们需要统计所有训练样本中的量化索引，并根据其频率对表征进行惩罚:

$$w_t^{\text{SIF}} = a/(a + N(q_t)), \tag{6}$$

其中，我们将 q 的计数记为 N（q），而 a 是平滑超参数，与原始 SIF [5] 类似。

然而，计数还可以采用其他策略，即用于惩罚的标准 w_t。我们设计了一种计数策略，在各组内分别搜索索引，充分利用组间相似性，同时避免稀疏性 [16]:

$$w_t^{\text{GP}} = 1/\textstyle\sum_{g=1}^{G} N_g(q_t) \tag{7}$$

其中， N_g 表示组 g 内的索引数量。为简化处理，我们还省略了平滑超参数 a。

另一方面，在第2.2节中，我们已经指出，在同一个实例中，相似表征的重要性可能较低。我们提出了考虑 VQ 索引全局概率（GP）的计数方法，同时也设计了另一种仅统计实例表征中索引的方法。两者唯一的区别在于计数范围从全局变为局部。我们将这种计数方法称为 LP。

最后，我们可以将 LP 和 GP 的权重相乘，同时利用局部概率和全局概率，并将其记为 BP：

$$w_t^{\mathrm{BP}} = w_t^{\mathrm{LP}} \cdot w_t^{\mathrm{GP}}$$

其中 w_t^{LP} 和 w_t^{GP} 分别是来自 LP 和 GP 的C权重。这种池化策略旨在降低那些在各个实例中冗余、且在整个训练集中频繁出现的量化索引的权重。

3. 实验结果

我们严格遵循 SUPERB 基准来衡量自监督语音模型的有效性 [2], , 但在无监督池化方法的评估上有所不同（第 3.2 节）。更多详情请参阅我们的代码库。[1]

3.1. 任务

我们选择了四种语音任务，以代表语音处理的不同方面。**关键词检测（KWS）**旨在自动检测预定义词汇表中的词语。我们使用 Google Speech Commands 数据集 V2 [18], , 其中包含 35 个关键词，包括数字和词语。**说话人识别（SID）**旨在根据给定的语句识别说话人。我们的基准测试聚焦于文本无关闭集说话人识别 [29]。我们使用 VoxCeleb1 数据集 [19] 及其提供的数据划分。**意图分类（IC）**从语音中提取语义信息，例如说话人的意图或情绪 [30]。我们使用流畅语音指令数据集 [20]，其中提供三种不同信息的标签：动作、对象和地点。**情感识别（ER）**是一项根据言语通道预测情绪类别的任务。我们采用 IEMOCAP [21] , 并选择 4 个情绪类别以遵循 [2]。我们使用 IEMOCAP 数据集中提供的 5 个会话进行 5 折交叉验证来训练模型。

3.2. 基准测试设计

我们在无监督设置下设计基准测试：不允许使用额外的可训练参数，以便公平比较无监督池化方法的有效性。我们通过测量测试数据集表征与话语级训练数据集表征之间的距离，为每个测试数据集表征寻找最近邻，从而在多种任务上评估这些表征。为衡量准确率，我们比较测试样本的标签与其最近邻训练样本的标签。这是基于对比损失的自监督表征中常用的设置 [22, 23]。我们使用近似最近邻方法 [24] 高效地计算距离。

3.3. 模型

我们采用了三种基于 wav2vec2.0、能够生成量化表征的不同自监督语音模型：基础版、大型版 [1] 和 XLS-R [25]。基础版与大型版的区别在于模型规模，而大型版与 XLS-R 的区别在于预训练数据集，前者使用朗读语音数据，后者使用来自各种来源的多语种语音语料库。我们选择这三种广泛使用的模型，以便评估池化技术在各种条件下的影响。

3.4. 基线方法

我们选择了四种无监督池化基线。**平均池化**（AP）是各种池化方法中最简单且应用最广泛的基线之一，

[1]https://github.com/IIP-Sogang/ 语音池化基准

表 1： 语音表示池化的无监督基准测试。报告通过最近邻方法（第 *3.2* 节）得出的测试准确率（%）。最佳准确率以粗体标示。

	方法	KWS	SID	IC	ER	AVG
base	AP	58.44	13.85	26.71	53.44	38.11
	SP	68.43	**15.63**	33.11	53.30	42.62
	白化	60.55	15.13	36.59	50.88	40.79
	软衰减	58.66	13.66	26.71	53.44	38.12
	VQ-Squash-AND	58.46	14.82	26.97	52.76	38.26
	VQ-Squash-OR	55.83	14.12	24.97	52.00	36.73
	VQ-AllSquash-AND	65.16	14.80	32.96	53.50	41.60
	VQ-AllSquash-OR	65.92	12.99	37.81	52.61	42.33
	VQ-SIF	59.62	13.33	36.54	54.28	40.94
	VQ-LP	**74.48**	15.53	48.11	**55.48**	**48.40**
	VQ-GP	67.98	12.91	41.55	54.11	44.14
	VQ-BP	73.48	12.94	**50.99**	55.13	48.14
large	AP	71.87	17.62	29.95	54.37	43.45
	SP	77.09	19.15	36.91	56.77	47.48
	白化	67.44	18.75	36.75	55.45	44.60
	软衰减	72.02	17.37	30.21	54.37	43.49
	VQ-Squash-AND	70.21	17.05	31.43	54.12	43.20
	VQ-Squash-OR	66.31	17.40	29.21	54.27	41.80
	VQ-AllSquash-AND	76.75	18.77	39.12	55.34	47.50
	VQ-AllSquash-OR	75.75	16.64	41.76	56.04	47.55
	VQ-SIF	73.09	17.60	32.64	54.59	44.48
	VQ-LP	**83.07**	**19.42**	55.00	**57.29**	**53.69**
	VQ-GP	76.57	15.89	51.86	55.28	49.90
	VQ-BP	81.00	14.71	**58.53**	54.28	52.13
XLS-R	AP	68.42	11.62	26.13	55.33	40.38
	SP	77.26	13.42	36.70	56.54	45.98
	白化	60.72	**15.22**	30.11	50.69	39.18
	软衰减	68.35	12.02	26.13	55.33	40.46
	VQ-Squash-AND	70.80	12.06	34.27	55.35	43.12
	VQ-Squash-OR	73.31	12.46	41.79	55.55	45.78
	VQ-AllSquash-AND	75.97	12.82	44.53	56.72	47.51
	VQ-AllSquash-OR	75.37	12.02	40.36	55.12	45.72
	VQ-SIF	80.45	11.89	46.08	57.46	48.97
	VQ-LP	82.38	12.71	49.14	**58.84**	50.77
	VQ-GP	82.87	12.00	52.49	57.74	51.27
	VQ-BP	**86.30**	11.71	**52.89**	57.60	**52.12**

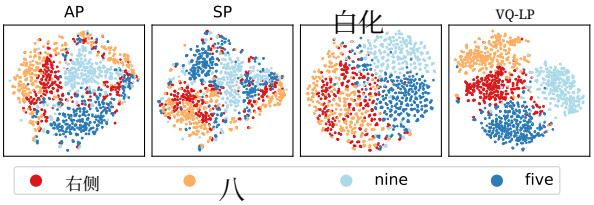

图 2： *Google Speech Commands V2* 数据集池化特征的 *t-SNE* 可视化。每个点代表一个数据实例，颜色表示四个不同的类别。

之一 [2]。**统计池化**（SP）在平均池化的基础上，不仅利用一阶统计量，还利用二阶统计量，即均值和标准差 [9,10]。**白化**对表征应用白化变换，使其协方差变为单位矩阵 [6,7]。**软衰减**是当前句子嵌入领域最先进的方法，它通过应用非线性函数，使表征的奇异值分布更加均匀 [8]。

3.5. 结果

表 1 展示了无监督基准测试结果。与平均池化（AP）相比，我们发现，采用 AND 和 OR 进行压缩分别使 XLS-R 的整体性能提升了 2.74% 和 5.4%，而其他模型则没有显著提升。此外，在 wav2vec 2.0 的所有变体中，AllSquash 策略的表现优于 Squash，并且超过了平均池化（AP）。另外，VQ-LP 在基础版和大型版中取得了最高的平均准确率，而 VQ-BP 在每种骨干网络上的意图分类（IC）任务中都表现出色，准确率是平均池化（AP）的两倍。

表 2： 与监督方法的性能比较。报告验证集所保留的最佳模型的测试准确率（%）。最佳准确率以粗体显示。

	方法	KWS	SID	IC	ER	AVG
base	AP	94.89	65.36	91.88	62.59	78.68
	SP	95.15	64.72	97.39	55.66	79.18
	VQ-LP	95.49	64.82	95.70	62.73	79.69
	SAP	95.15	40.86	97.39	55.66	72.27
	ASP	95.70	36.52	97.47	59.34	72.26
	VAP	**96.74**	**68.50**	**99.50**	**65.50**	**82.56**
large	AP	96.05	**79.23**	95.07	65.47	83.96
	SP	96.23	77.71	96.73	67.02	84.42
	VQ-LP	96.32	78.12	97.76	66.58	84.70
	SAP	96.26	58.48	96.78	57.78	77.33
	ASP	96.42	46.36	98.55	61.94	75.82
	VAP	**97.14**	77.08	**99.58**	**67.42**	**85.30**
XLS-R	AP	97.32	67.85	91.30	71.77	82.06
	SP	97.51	60.89	94.12	**72.07**	81.15
	VQ-LP	97.59	**68.13**	94.38	70.67	**82.69**
	SAP	**97.63**	61.73	98.86	64.83	80.76
	ASP	96.69	13.48	99.02	61.08	67.57
	VAP	97.49	61.71	**99.63**	71.15	82.49

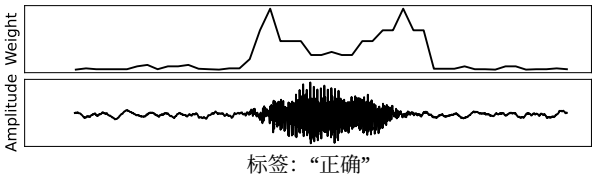

图 3： *VQ-LP* 产生的权重可视化。两个明显的峰值表明存在“*r*”和“*t*”音，而对应元音的区域中数值则有所下降。

我们还使用 t 分布随机邻域嵌入（t-SNE） [31] 来可视化池化表示，如图 2 所示。观察结果表明，基于VQ的方法有助于区分元音发音相似的关键词表征。总体而言，我们的结果证明，基于VQ的池化方法能够有效捕捉语音表征的时间重要性。

4. 分析

4.1. 与参数化方法的比较

本节比较了我们的方法与参数化方法的效果。我们选取了说话人识别中常用的参数化池化方法：自注意力池化（SAP） [32], 注意力统计池化（ASP）[11], 和基于向量的注意力池化（VAP） [13]。

为在监督设置下比较所有方法，我们使用了 Adam优化器 [33]，默认学习率为 0.001，训练 200 个 epoch；例外情况是，KWS 的学习率设为 0.01，SID 训练 50 个 epoch。请注意，参数化方法具有额外的可训练参数，可针对下游任务进行拟合，因此这种比较对它们更有利。

尽管使用量化表征的方法在无监督实验中带来了显著提升，但表 2 中并未观察到同等程度的提升。如何——

不过，VQ-LocalProb 的表现与参数化方法相当，在 XLS-R 上甚至取得了高于最佳参数化方法 VAP 的平均准确率。这些结果表明，无监督池化方法可能进一步超越监督式池化方法，

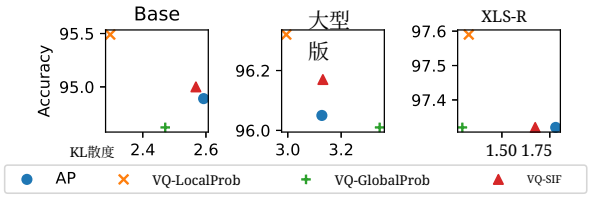

图 4： 不同池化方法权重的比较。*KL* 散度衡量 *SAP* 与其他方法所得权重之间的相似度，并将 *SAP* 视为预言机。我们观察到，权重越接近 *SAP*，下游性能通常越好。

表 3： 基础版上采用平均池化时，不同最近邻数量 *(K)* 对应的测试准确率 *(%)*。

Task \ K	1	3	4	5	6	7	8	9
KWS	58.44	60.42	61.88	62.48	62.58	62.85	63.22	63.12
SID	13.85	13.59	13.38	13.17	12.94	12.63	12.14	11.93
IC	26.71	27.58	27.74	27.42	28.16	28.16	28.08	28.05
ER	53.44	56.68	57.54	58.45	59.30	59.30	60.11	59.51

值得在未来的研究中继续探索。

4.2. 池化权重分析

我们通过将不同策略的权重与 SAP 进行比较，分析了所提出池化方法的行为。我们使用 KL 散度 $D_{\mathrm{KL}}(P\|Q)$，分别将 SAP 和其他方法产生的权重放入 P 和 Q。图 4 显示，我们提出的方法通常比 AP 更接近 SAP。我们还发现，准确率与 KL 散度呈负相关，这证实了与 SAP 的相似度越高，性能越好。此外，我们在图 3 中展示了来自 Speech Commands V2 的一段语音样本及其由 VQ-LP 方法产生的池化权重；我们的方法能够检测语音，同时抑制含有环境噪声的区域。

4.3. 探索不同的评估方案

由于第 3.2 节中的基准测试设计仅使用最近邻的信息，我们进一步探讨了考虑一组邻居的情形。我们通过实验检验，增加最近邻的数量并通过投票确定最终标签，是否能够提升下游性能。如果出现平票，则采用距离最近样本的标签，因此我们不考虑 2 个最近邻的情形。表 3 展示了不同最近邻数量下的性能差异。除 SID 外，所有任务的性能均有所提升；SID 任务包含的类别相对较多。这表明，当前的基准测试设置在更贴近现实的应用场景中仍有提升空间。

5. 结论

本文设计了一种无监督但有效的方法，用于池化自监督语音模型提取的可变长度语音特征。在各种语音处理任务上的下游性能以及自监督表征验证了我们基于VQ的方法的有效性。此外，我们将这种无监督池化方法与基于注意力机制的有监督池化方法进行比较，以了解其内部工作机制。未来应进一步研究VQ方法在更广泛的通用语音模型中的潜力，这些模型不包含内置的VQ模块，例如融入HuBERT [34]的理念。

6. 参考文献

[1] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：一种用于语音表征自监督学习的框架”，神经信息处理系统进展，第33卷，第12 449–12 460页，2020年。[2] S. Yang、P. Chi、Y. Chuang、C. J. Lai 等，“SUPERB：语音处理通用性能基准测试”，载于*INTERSPEECH*论文集，2021年。[3] J. D. M.-W. C. Kenton 和 L. K. Toutanova，“BERT：用于语言理解的深度双向变换器预训练”，载于*NAACL-HLT*论文集，2019年，第4171–4186页。[4] J. Wieting、M. Bansal、K. Gimpel 和 K. Livescu，“迈向通用释义句嵌入”，载于*ICLR*论文集，2016年。[5] S. Arora、Y. Liang 和 T. Ma，“一种简单却难以超越的句子嵌入基线”，载于*ICLR*论文集，2017年。[6] J. Huang、D. Tang、W. Zhong 和 S. e. a. Lu，“WhiteningBERT：一种简单的无监督句嵌入方法”，载于*EMNLP*论文集，2021年，第238–244页。

[7] J. Su, J. Cao, W. Liu, 和 Y. Ou，“白化句子表示 - 表示以获得更好的语义和更快的检索”，*arXiv* 预印本 *arXiv:2103.15316*，2021年。

[8] H. Yan、L. Gui、W. Li 和 Y. He，“通过奇异值变换解决 Transformer 中的令牌均匀性问题”，载于人工智能不确定性会议论文集。PMLR，2022年，第 2181–2191 页。[9] D. Snyder、P. Ghahremani、D. Povey、D. Garcia-Romero、Y. Carmiel 和 S. Khudanpur，“用于端到端说话人验证的深度神经网络说话人嵌入”，载于口语语言技术研讨会（*SLT*）论文集。IEEE，2016年，第 165–170 页。[10] D. Snyder、D. Garcia-Romero、D. Povey 和 S. Khudanpur，“用于文本无关说话人验证的深度神经网络嵌入”。载于*INTERSPEECH* 论文集，2017年，第 999–1003 页。[11] K. Okabe、T. Koshinaka 和 K. Shinoda，“用于深度说话人嵌入的注意力统计池化”，载于*INTERSPEECH* 论文集，2018年。[12] P. Safari、M. India 和 J. Hernando，“用于说话人识别的自注意力编码与池化”，载于*INTERSPEECH* 论文集，2020年，第 941–945 页。[13] Y. Wu、C. Guo、H. Gao、X. Hou 和 J. Xu，“用于文本无关说话人验证的基于向量的注意力池化”。载于*INTERSPEECH* 论文集，2020年，第 936–940 页。

[14] F. K. Soong, A. E. Rosenberg, B.-H. Juang, 和 L. R. Rabiner, “报告：一种用于说话人识别的矢量量化方法”，*AT&T* 技术期刊，第 66 卷，第 2 期，第 14–26 页，1987 年。

[15] H. Jegou、M. Douze 和 C. Schmid，“用于最近邻搜索的乘积量化”，*IEEE* 模式分析与机器智能汇刊，第 33 卷，第 1 期，第 117–128 页，2010 年。[16] A. Baevski 、 S. Schneider 和 M. Auli，“vq-wav2vec：离散语音表征的自监督学习”，载于*ICLR* 论文集，2020年。

[17] K. Choi, Yeo 和 E.《揭开 wav2vec 特征编码器的黑箱》，*arXiv* 预印本 *arXiv:2210.15386*，2022年。

[18] P. Warden，《语音命令：一个用于有限词汇语音识别的数据集》，*arXiv* 预印本 *arXiv:1804.03209*，2018年。[19] A. Nagrani、J. S. Chung 和 A. Zisserman，《VoxCeleb：大型说话人识别数据集》，载于 *INTERSPEECH* 论文集。ISCA，2017年，第2616–2620页。

[20] L. Lugosch、M. Ravanelli、P. Ignoto、V. S. Tomar 和 Y. Ben-gio，《用于端到端口语理解的语音模型预训练》，载于*INTERSPEECH* 论文集，2019年。

[21] C. Busso、M. Bulut、C.-C. Lee、A. Kazemzadeh、E. Mower、S. Kim、J. N. Chang、S. Lee 和 S. S. Narayanan，“IEMOCAP：交互式情感双人动作捕捉数据库”，《语言资源与评估》，第 42 卷，第 335–359 页，2008年。

[22] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton，“一种用于视觉表征对比学习的简单框架”，载于《国际机器学习会议论文集》。PMLR，2020年，第 1597–1607 页。[23] K. Choi 和 S. Lee，“通过正则化应对基于互信息的损失的不稳定性”，载于《人工智能不确定性会议论文集》。PMLR，2022年，第 411–421 页。[24] E. Bernhardsson，“Annoy：针对内存使用以及磁盘加载/保存进行优化的 C++/Python 近似最近邻”，*GitHub https://github.com/spotify/annoy*，2017年。[25] A. Babu、C. Wang、A. Tjandra、K. Lakhotia、Q. Xu、N. Goyal、K. Singh、P. von Platen、Y. Saraf、J. Pino、A. Baevski、A. Conneau 和 M. Auli，“XLS-R：大规模自监督跨语言语音表征学习”，载于*INTERSPEECH*论文集，2022年，第 2278–2282 页。[26] T. Zenkel、R. Sanabria、F. Metze、J. Niehues、M. Sperber、S. St¨uker 和 A. Waibel，“CTC声学模型解码策略的比较”，*arXiv*预印本 *arXiv:1708.04469*，2017年。[27] K. Sparck Jones，“词项特异性的统计解释及其在检索中的应用”，《文献学杂志》，第 28 卷，第 1 期，第 11–21 页，1972年。[28] T. Mikolov、W.-t. Yih 和 G. Zweig，“连续空间词表示中的语言规律”，载于计算语言学协会北美分会人类语言技术会议论文集，2013年，第 746–751 页。[29] M. M. Kabir、M. F. Mridha、J. Shin、I. Jahan 和 A. Q. Ohi，“说话人识别综述：基础理论、识别方法与机遇”，*IEEE Access*，第 9 卷，第 79 236–79 263 页，2021年。[30] S. Seo、D. Kwak 和 B. Lee，“将预训练网络与连续词元接口集成，用于端到端口语理解”，载于国际声学、语音与信号处理会议（*ICASSP*）论文集。IEEE，2022年，第 7152–7156 页。[31] G. E. Hinton 和 S. Roweis，“随机邻居嵌入”，载于*NeurIPS*，第 15 卷，2002年。[32] W. Cai、J. Chen 和 M. Li，“探索端到端说话人与语言识别系统中的编码层和损失函数”，载于*Odyssey*说话人与语言识别研讨会论文集，2018年，第 74–81 页。[33] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于*ICLR*论文集，Y. Bengio 和 Y. LeCun 编，2015年。[34] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过隐藏单元掩码预测进行的自监督语音表征学习”，*IEEE/ACM*音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021年。