arXiv:2303.00733v1 [eess.AS] 1 Mar 2023

SPEECHPR 语音分类任务的提示调优 S

Kai-Wei Chang[1⋆] *Yu-Kai Wang*[2⋆] *Hua Shen*[3] *Iu-thing Kang*[4] *Wei-ChengTseng*[1] *Shang-Wen Li*[5] *Hung-yi Lee*[1]

[1]国立台湾大学通信工程研究所，台湾[1,2,4]国立台湾大学，台湾，[3]宾夕法尼亚州立大学，[5]Meta AI

摘要

提示调优是一种技术，通过调优一小组参数来引导预训练语言模型（LM）直接生成下游任务的输出。近年来，提示调优在自然语言处理（NLP）和语音处理领域均展现出存储和计算效率方面的优势。这些优势也表明，提示调优可作为一种候选方法，以统一的方式让预训练语言模型服务于多个任务。在语音处理方面，SpeechPrompt 在若干语音分类任务上展现出较高的参数效率和具有竞争力的性能。然而，SpeechPrompt 是否能够服务于大量任务仍未得到解答。在本工作中，我们提出了 SpeechPrompt v2，这是一个提示调优框架，能够执行多种多样的语音分类任务，涵盖多种语言和韵律相关任务。实验结果表明，在统一框架中，SpeechPrompt v2 仅使用不到 0.15M 个可训练参数，即可达到与先前工作相当的性能。SpeechPrompt 项目网站为 https://ga642381.github.io/SpeechPrompt。

索引词— 提示调优、自监督学习、口语语言模型、语音分类

1. 引言

近年来，自监督预训练模型已成为语音处理的重要组成部分。通过利用从大规模无标注语料库中学习到的结构知识，这些模型能够生成有益于各种语音处理任务的通用表征 [1]。在将自监督语音模型用于所关注的下游任务时，一种典型做法是遵循**“预训练、微调”范式**。在这一范式下，特定任务的下游模型构建于预训练的自监督语音模型之上。随后，使用下游任务的目标函数对两个模型进行联合微调。已有多项研究表明，这一范式能够在各种语音处理任务上取得最先进的性能。然而，这一范式实际上

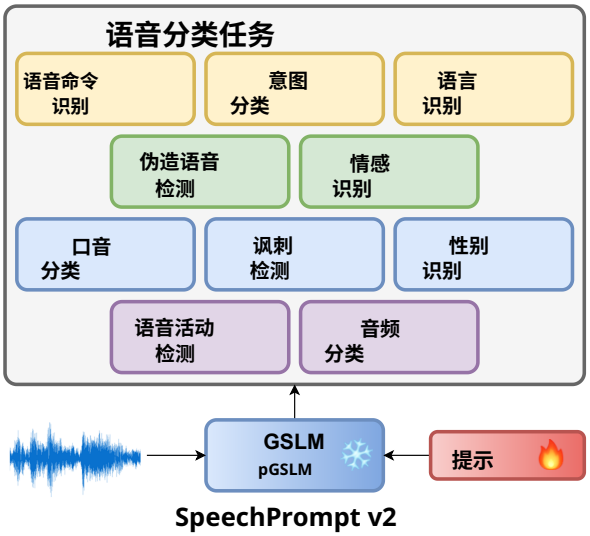


图 1：提出的 SpeechPrompt v2。它展现出出色的通用性，在各种语音分类任务中均表现出竞争力，且只需更新少量参数。预训练的口语语言模型保持冻结状态，只有少量提示向量可训练。

要求为每个任务精心设计下游模型和损失函数，从而不断增加人工负担。此外，当下游任务数量扩展时，它还需要大量的计算和存储资源。

另一方面，研究人员探索了**提示范式** [2]，将其作为利用预训练模型的替代方法。提示是指寻找特定于任务的模板或指令，以引导预训练语言模型，而无需修改其架构和参数。提示方法可以通过掩码预测或生成方式服务于下游任务。例如，在情感分类中，可以将输入句子 $\langle S \rangle$ 填入模板：“$\langle S \rangle$。它很。”，然后输入预训练语言模型。随后，标签映射器（通常是人工设计的映射函数）会将语言模型的输出（如 great、terrible）转换为情感类别（positive、negative）。这种方法使我们能够判断 $\langle S \rangle$的情感。

提示相比典型的“预

⋆ 前两位作者贡献相同。

训练、微调”范式有诸多优势。例如，我们不再需要为每项任务设计一个下游模型。所有任务都可以在统一框架中处理，从而提高计算和存储效率。在固定预训练语言模型的前提下，人们提出了一种高效的提示方法——**提示调优**，它通过更新少量连续提示向量来实现。例如，前缀调优 [3] 会执行深度提示调优，将可训练提示拼接到语言模型各 Transformer 层的输入中。

SpeechPrompt [4]将提示调优应用于生成式口语语言模型（GSLM） [5]，以执行语音分类和序列生成任务。然而，它仅展示了在有限的语音处理任务和数据集上采用提示范式的可行性。为此，我们希望进一步考察提示范式在广泛语音分类任务上的泛化能力。我们提出了**SpeechPrompt v2**，这是一种提示调优框架，能够执行各种语音分类任务，包括内容相关和韵律相关任务，并涵盖多种语言。我们的贡献主要有三点:

- 我们提出了*SpeechPrompt v2*，这是一种提示调优框架，能够执行范围广泛的语音分类任务；在采用统一模型架构和有限可训练参数的情况下，其性能可与先前研究相媲美。
- 我们引入了一组语音分类任务，用于评估自监督口语生成模型。这些任务涵盖多种语言中的内容相关和韵律相关任务。
- 我们引入了一种可学习 verbalizer，能够稳定提升口语语言模型采用提示范式时的性能。此外，我们还进行了分析，以验证其必要性。

2. 相关工作

2.1. 生成式语音语言模型

近期，人们提出了生成式**口语语言模型**，用于建模各种语音特征，并在生成新颖语音方面展现出有竞争力的性能。**生成式口语语言模型（GSLM）** [5] 在离散单元上执行生成式语言建模，这些单元充当伪文本，承载由自监督学习语音模型编码的语音信息。此外，**韵律感知生成式口语语言模型（pGSLM）** [6] 通过对语音韵律进行建模，进一步拓展了 GSLM 的能力。研究者提出了一种多流 Transformer，用于同时建模语音信号的离散单元、时长和音高信息。这些口语语言模型以生成方式运行，可视为 GPT 的语音版本 [6], ，因此为在语音处理领域进行提示提供了可能。

2.2. 语音处理中的提示与重编程

提示最初源自自然语言处理领域，如今也已在语音处理领域得到探索。WAVPROMPT [7]采用 GPT-2 作为骨干语言模型，并将音频编码到提示中，以执行口语理解（SLU）任务。SpeechPrompt [4] 在 GSLM 上进行提示调优，以完成语音分类和序列生成任务。研究者引入了一种启发式频率标签映射器，将 GSLM 的词汇单元映射到下游任务的类别。然而，基于频率的标签映射器并未显著优于随机映射方法。为此，我们在 SpeechPrompt v2 中提出了一种可学习 verbalizer，以提升性能。

利用预训练模型能力完成不同任务的另一种方法是**模型重编程**。在 [8, 9],中，首先使用任务特定函数对输入数据（目标域）进行转换。随后，预训练声学模型能够为重编程后的数据生成标签。这些标签（源域）随后通过映射函数映射到下游任务（目标域）的类别。该函数在提示方法中发挥着与标签映射器相同的作用，在重编程文献中通常采用随机映射。

3. 方法

我们提出了 SpeechPrompt v2，这是一种提示框架，面向服务于多种下游语音分类任务的口语语言模型。在一般的提示调优方法和 SpeechPrompt v2 中，预训练语言模型始终保持固定。给定一个下游任务和预训练口语语言模型 $\mathcal{M}$，我们学习任务特定提示向量 $\mathcal{P}$。通过基于输入 $\boldsymbol{x}$ 和提示向量 $\mathcal{P}$对语言模型 $\mathcal{M}$进行条件化，我们可以生成语言模型的输出，随后使用标签映射器 v 将其映射为任务标签，从而得到预测:

$$\hat{\boldsymbol{y}} = v(\mathcal{M}(\mathcal{P}; \boldsymbol{x})). \tag{1}$$

在 SpeechPrompt v2 中，我们采用 GSLM 和 pGSLM 作为骨干语言模型。下文将介绍 SpeechPrompt v2 的具体实现细节，以便将通用提示技术引入语音领域。

3.1. 提示调优

在 GSLM 和 pGSLM 中，首先通过自监督学习模型和 K-means量化器将一段话语编码为离散单元 $\boldsymbol{u}$。
GSLM 与 pGSLM 的区别在于，pGSLM 还会从给定话语中提取时长 $\boldsymbol{d}$ 和音高信息 $\boldsymbol{f}$（对数基频）。随后，提示向量 $\boldsymbol{p}^I$ 会在输入处进行拼接

$$\boldsymbol{x} = \begin{cases} [\boldsymbol{p}^I, \boldsymbol{e}(\boldsymbol{u})] & \text{for GSLM} \\ [\boldsymbol{p}^I, \boldsymbol{e}(\boldsymbol{u}) + \boldsymbol{e}(\boldsymbol{d}) + \boldsymbol{e}(\boldsymbol{f})] & \text{for pGSLM} \end{cases}, \tag{2}$$

其中 $\boldsymbol{p}^I$ 是 $\mathcal{P}$中的可训练向量，而 $\boldsymbol{e}(\cdot)$ 表示口语语言模型的输入嵌入。

除了在输入端拼接提示外，我们还采用深度提示调优 [3, 4] 来增强提示的指令能力。给定以嵌入表示 $\boldsymbol{x}^j = [\boldsymbol{x}_1^j, \boldsymbol{x}_2^j, \cdots, \boldsymbol{x}_T^j]$ 为输入的 j^{th} Transformer 层，我们将提示向量 $\boldsymbol{p}^K$ 和 $\boldsymbol{p}^V$ 拼接起来，以调节注意力函数 $Attn(Q, K, V)$ 中的键和值：

$$K = Concat(\boldsymbol{p}^K, \boldsymbol{x}_{l+1:T}^j)W^K, \tag{3}$$

$$V = Concat(\boldsymbol{p}^V, \boldsymbol{x}_{l+1:T}^j)W^V, \tag{4}$$

其中，$\boldsymbol{p}^K$、$\boldsymbol{p}^V$ 是 $\mathcal{P}$ 中的可训练提示，l 是提示长度。

3.2. 可学习 verbalizer

为了将语言模型的输出与下游任务的类别关联起来，通用提示范式会采用标签映射器 v，而 v 则将语言模型的词汇表映射到任务类别。例如，在情感分类中，“great”一词会映射到正向情感。SpeechPrompt [4] 采用基于统计频率的映射作为标签映射器，无需进行模型估计。然而，该算法存在信息丢失问题（图 2），且其性能并不总是优于随机映射方法。在 SpeechPrompt v2 中，我们提出使用可学习的线性模型来实现标签映射器 v。v 以口语语言模型在词汇表上的输出分布为输入，并将该分布映射到下游任务的类别。因此，可训练提示向量 $\mathcal{P}$ 与标签映射器 v 通过损失函数 $\mathcal{L}$联合优化：

$$\mathcal{P}, v = \arg\min_{\mathcal{P}, v} \mathcal{L}(v(\mathcal{M}(\mathcal{P}; \boldsymbol{x})), \boldsymbol{y}), \tag{5}$$

其中 $\boldsymbol{y}$ 表示下游任务的真实标签。在 SpeechPro mpt v2 中，我们对每项任务都使用交叉熵。

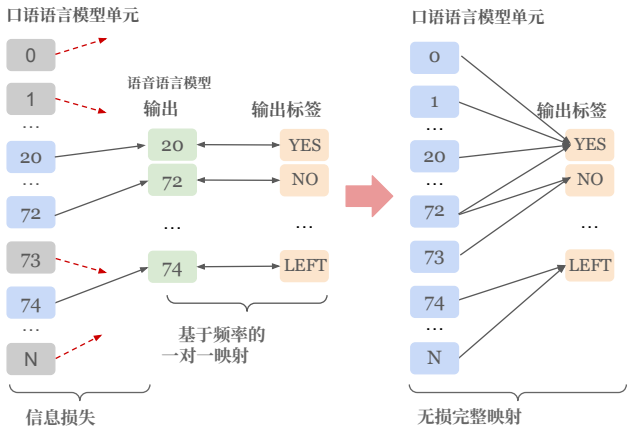
4. 实验设置

4.1. 任务与数据集

本文在来自 14 个数据集的 10 项语音分类任务上评估 SpeechPrompt v2，这些数据集具有不同的语言和语音特征。下面我们简要介绍这些任务 [1]。

语音命令识别，SCR：SCR旨在识别给定话语中出现的关键词。我们

[1]本实验仅旨在依据既定基准和标签评估模型能力。这些既定标签并未全面反映性别和情绪的多样性。根据音频预测二元性别存在误判他人性别的风险，也无法涵盖所有性别认同，包括非二元性别者。此外，表面上呈现的情绪并非普遍一致，而是受文化影响，不同的人可能会以不同方式通过音频表达情绪。



(A) 频率映射 verbalizer　　（B）可学习标签映射器

图 2：频率映射标签映射器（A）与 Speech-Prompt v2 中提出的可学习标签映射器（B）的比较。

采用了 Google Speech Commands [20] 以及多个不同语言的低资源数据集，包括 Grabo Speech Commands [21], 立陶宛语语音命令 [22],和阿拉伯语语音命令 [23]。

意图分类，IC：IC 旨在将话语划分到预定义的类别中，以确定说话者的意图。我们使用了 Fluent Speech Commands 数据集 [24], , 其中每条话语都有三个标签：动作、对象和位置。

语言识别，LID：LID 旨在识别话语所使用的语言。我们使用了包含六种不同语言的 Voxforge 数据集 [25]来执行该任务。

伪造语音检测，FSD：FSD 旨在区分真实语音与合成语音。我们使用了 ASVspoof 数据集 [26], 中的逻辑访问（LA）部分，该部分包含真实语音和伪造语音。

情感识别，ER：ER 旨在识别给定话语所表达的情绪。本研究采用了广泛使用的情感识别数据集 IEMOCAP [27]。

口音分类，AcC：AcC 旨在对同一种语言的不同口音进行分类。我们使用了 AccentDB 数据集 [14], , 其中包含 4 种印度英语口音、4 种母语英语口音和 1 种印度大都市英语口音。

讽刺检测，SD：SD 旨在检测一段话语是否包含讽刺。我们采用了两个数据集，即 MUS- tARD [15] 和 MUStARD++ [16]。后者是前者的扩展。

性别识别，GID：GID 旨在将一段话语分类为不同性别。我们采用了 VoxCeleb1 数据集 [28]，并根据所提供的说话人信- 息获得标签。

语音活动检测，VAD：VAD 旨在检测输入音频是否包含语音或背景噪声。我们使用 Google Speech Commands v2 [20] 作为语音数据，使用 Freesound 数据集 [29] 作为背景噪声数据，并将两者混合

表1：SpeechPrompt v2 在多种语音分类任务上的表现。“+”表示采用可学习 verbalizer。对于每项任务，我们还列出了以“全监督学习”或“预训练、微调”范式取得的先前最先进水平。这些工作采用相同的数据集，作为性能上限。括号中的数字表示 SpeechPrompt 与最先进水平之间的性能差异。为与先前工作进行比较，结果以 FSD 的等错误率（EER）、SD 和 GID 的 F1 分数（F1），以及其余任务的准确率（ACC）报告。

Task	指标	数据集	语言	#类别		GSLM	GSLM+	pGSLM	pGSLM+
SCR	准确率 (↑)	Google SC v1	En	12	98.6 [10]	94.5	94.6	94.3	**94.7** (-3.9)
		Grabo SC	Du	36	98.9 [11]	92.4	**92.7** (-6.2)	17.5	19.6
		立陶宛语 SC	Lt	15	91.8 [9]	93.2	**95.5** (+3.7)	90.9	79.5
		阿拉伯语 SC	Ar	16	98.9 [9]	99.7	**100.0** (+1.1)	85.6	92.6
IC	准确率 (↑)	Fluent SC	En	24	99.7 [12]	97.2	97.3	98.1	**98.2** (-1.5)
LID	准确率 (↑)	Voxforge	英语、西班牙语、法语 De, Ru, It	6	99.8 [13]	90.9	**94.2** (-5.6)	81.8	80.4
FSD	等错误率 (↓)	ASVspoof	En	2	2.5 [13]	18.5	13.5	**13.1** (+10.6)	18.3
ER	准确率 (↑)	IEMOCAP	En	4	79.2 [13]	42.1	44.3	49.9	**50.2** (-29)
AcC	准确率 (↑)	AccentDB	En	9	99.5 [14]	78.9	83.4	86.5	**87.1** (-12.4)
SD	F1 (↑)	MUStARD	En	2	64.6 [15]	55.0	77.8	74.4	**78.7** (+13.1)
		MUStARD++	En	2	65.2 [16]	74.0	**75.2** (+10)	52.7	58.2
GID	F1 (↑)	VoxCeleb1	En	2	98.3 [17]	86.2	87.3	**91.6** (-6.7)	86.2
VAD	准确率 (↑)	Google SC v2 & Freesound	En	2	98.8 [18]	96.6	96.9	**98.3** (-0.5)	98.1
AuC	准确率 (↑)	ESC-50	✖	50	97.0 [19]	9.0	**37.5** (-59.5)	20.3	27.0

以评估语音活动检测。

音频分类，AuC：AuC旨在对人类活动和自然界产生的不同环境声音进行分类。我们使用了包含 50 个不同类别的 ESC-50 [30] 完成此任务。

4.2. 实现细节

预训练语言模型对于 GSLM 和 pGSLM 上的提示调优，我们使用具有 100 个簇的预训练 HuBERT [31]，以及与之对应的预训练 GSLM 和 pGSLM。预训练的语音自监督学习模型、量化器和口语语言模型可在*fairseq*[2]上获取。

提示和标签映射器为简化流水线，我们不会针对每项任务分别搜索超参数。对于每项任务，我们使用长度为 $l = 5$ 的提示，包含 0.128M 个参数。可学习的标签映射器是一个简单的线性模型，将离散单元映射到各任务的类别，且所含参数数量可以忽略不计。总体而言，每项任务中的可训练参数不到语音语言模型参数总量（~150M）的 0.1%。对于基线方法 SpeechPrompt（结合 GSLM 和 pGSLM），我们使用随机映射标签映射器，而非基于频率的标签映射器。在初步研究中，我们发现后者的表现并未始终优于前者。

[2]https://github.com/facebookresearch/fairseq/tree/main/examples/textless_自然语言处理

5. 结果

5.1. 语音分类任务

表1展示了 SpeechPrompt v2 在多种语音分类任务上的结果。SpeechPrompt v2 的表现可分为三类：

(1) 超越最先进水平：SpeechPrompt v2 在立陶宛语 SCR、阿拉伯语语音分类和讽刺检测（SD）任务上取得了新的最先进成果。该模型在不同语言间具有良好的泛化能力，在立陶宛语和阿拉伯语等低资源语言中表现尤为出色 [9]。借助强大的预训练口语语言模型，SpeechPrompt v2 在讽刺检测任务上也达到了最先进的性能。

(2) 与最先进水平相比具有竞争力：SpeechPrompt v2 在 Google语音分类、Grabo语音分类、意图分类（IC）、语言识别（LID）、性别分类（GID）和语音活动检测（VAD）任务上的表现与最先进水平相当。尽管在这些任务上的性能略低于当前最先进的系统，SpeechPrompt v2 仍具有竞争力。

(3)表现不及最先进水平：SpeechPrompt v2 在伪造语音检测（FSD）、情感识别（ER）、口音分类（AcC）和音频分类（AuC）任务上的表现不及最先进水平。实验表明，SpeechPrompt v2 在非语音

以及具有多样变化的语音上表现不佳。

实验还证明了在 SpeechPrompt 中使用可学习 verbalizer 的优势。借助可学习 verbalizer，采用 GSLM 的 SpeechPrompt 在大多数任务中都取得了稳定的性能提升。不过，对于立陶宛语 SCR 和 GID 等特定任务，在 pGSLM 中引入可学习 verbalizer 后，性能反而有所下降，这可能是提示调优不稳定所致。在下一节中，我们将深入探讨为何在 SpeechPrompt 框架中使用可学习 verbalizer 至关重要。

不过，实验结果也表明 SpeechPrompt 存在一些局限性。首先，在处理非语音数据或具有较大差异的语音时，其性能会下降。这可能是因为口语语言模型是在仅含英语语音的语料库上预训练的。其次，提示调优的性能存在一定波动，例如 pGSLM 在 Grabo SC 上的表现不佳。我们推测，这些不一致是由提示调优训练的不稳定性导致的。为保持框架简单且统一，我们没有对表1所列的所有实验调整任何超参数，包括提示长度、学习率等。因此，如何稳定提示调优仍有待未来研究。解决这一问题的可行方法包括引入超网络 [32]和提示初始化 [3]。

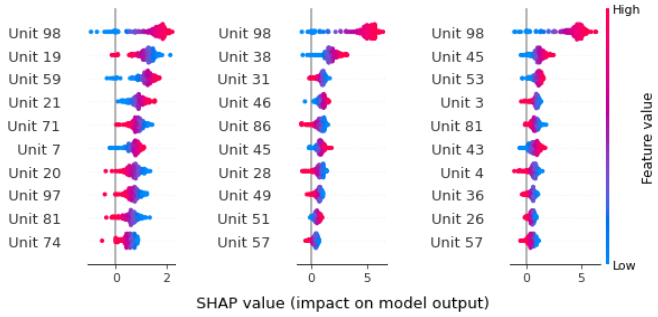
总体而言，SpeechPrompt 在广泛的语音分类任务上展现出了相当甚至达到最先进水平的性能。尽管在某些任务上性能有所下降，但得益于其提示范式，SpeechPrompt 具备参数效率高、统一前馈流程等优势。这些特性使 Speech-Prompt v2 成为当前系统的一个有前景的替代方案，尤其适用于优先考虑效率和简洁性的应用。

5.2. 可学习 verbalizer 分析

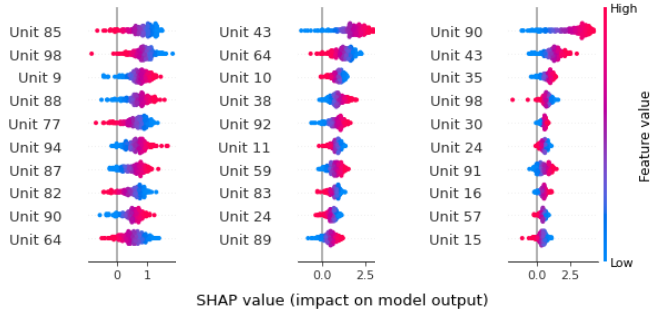
本节将进一步考察可学习 verbalizer 的行为，以了解其如何将离散单元映射为具有语义的任务标签。具体而言，我们采用 Shapley Additive ex-Planation [33]（SHAP）这一用于解释模型预测的常用方法。在 SHAP 中，**SHAP 值**用于衡量每个输入特征（此处指离散单元）对于模型所做单次预测的重要性。

图 3 展示了在 Google SC 数据集上进行三次不同运行后得到的 SCR 结果。我们列出了对“YES”类和“LEFT”类 [3] 贡献最大的前 10 个单元。在每张图中，横轴为 SHAP 值，纵轴按从上到下的顺序表示各单元的重要性。每个点代表对一个数据点的解释。点的颜色反映了各输入单元取值的大小。首先，我们观察到，即使属于同一类别，

[3]如需查看更多演示，请参阅我们的项目网站。https://ga642381.github.io/SpeechPrompt



(a) “YES”类



(b) “LEFT”类

图 3：可学习 verbalizer 在 Google Speech Command 数据集 [10] 上的 SHAP 分析。每个点代表一个数据点。

各单元的重要性也会因预测而异。可学习 verbalizer 依赖多个单元来确定输出，表明从单元到标签存在多对多映射行为。这验证了前人工作中可能存在的信息损失，如第 3.2 节所述。此外，不同运行中单元与标签之间的关系并不一致。不过，各次运行中存在共同的单元，且它们的贡献模式相似（例如，单元 98 对“YES”类的贡献模式相同）。最后，可学习 verbalizer 可能会利用一个单元来区分不同类别。例如，当单元 98 的值较低时，标签映射器更有可能预测为“LEFT”类而非“YES”类。总体而言，分析表明，与朴素随机映射和基于频率的映射相比，使用可学习 verbalizer 更有优势，也更适合我们的框架。

6. 结论

在这项工作中，我们提出了 SpeechPrompt v2，这是一种提示调优框架，能够执行多种语音分类任务。我们在 SpeechPrompt 中引入了可学习的 verbalizer，以增强分类能力。我们使用两种口语语言模型 GSLM 和 pGSLM 进行了实验。实验结果表明，SpeechPrompt v2 在不同的语音

分类任务中，在统一框架下实现了具有竞争力的性能和参数效率。我们还讨论了 SpeechPrompt 的局限性，这些讨论可为未来研究提供指引。总的来说，我们相信这项工作能够激励语音领域进一步探索提示范式在语音处理中的潜力。

7. 致谢

此处介绍的部分工作是在约翰斯·霍普金斯大学举办的2022年Jelinek纪念语音与语言技术暑期研讨会期间完成的，该研讨会由亚马逊、Mi- crosoft和谷歌提供不受限制的捐赠资助。

8. 参考文献

[1]A.Mohamed、H.-y.Lee、L.Borgholt、J. D.Havtorn、J.Edin、C.Igel、K.Kirchhoff、S.-W.Li、K.Livescu、L.Maaløe等，《自监督语音表征学习：综述》，*arXiv*预印本 *arXiv:2205.10643*，2022年。

[2] P.Liu、W.Yuan、J.Fu、Z.Jiang、H.Hayashi和G.Neubig，《预训练、提示与预测：自然语言处理领域提示方法的系统综述》，*arXiv*预印本 *arXiv:2107.13586*，2021年。

[3] X. L.Li 和 P.Liang，“前缀调优：优化用于生成的连续提示”，发表于*ACL/IJCNLP(1)*。2021 年，第 4582–4597 页，计算语言学协会。

[4]K.-W.Chang、W.-C.Tseng、S.-W.Li 和 H.-y.Lee，“SpeechPrompt：探索生成式口语语言模型在语音处理任务中的提示调优”，2022 年。

[5]K.Lakhotia、E.Kharitonov、W.-N.Hsu、Y.Adi、A.Polyak、B.Bolte、T.-A.Nguyen、J.Copet、A.Baevski、A.Mohamed 等，“基于原始音频的生成式口语语言建模”，计算语言学协会汇刊，第 9 卷，第 1336–1354 页，2021 年。

[6]E.Kharitonov、A.Lee、A.Polyak、Y.Adi、J.Copet、K.Lakhotia、T. A.Nguyen、M.Rivi`ere、A.Mohamed、E.Dupoux 和 W.Hsu，“无文本韵律感知生成式口语语言建模”，载于*ACL (1)*，2022 年，第 8666–8681 页。

[7]H.Gao, J.Ni, K.Qian, Y.Zhang, S.Chang 和 M.Hasegawa-Johnson，《Wavprompt：利用冻结语言模型实现少样本口语语言理解》，载于*INTERSPEECH*。2022 年，第 2738–2742 页，ISCA。

[8] C. H.Yang、Y.Tsai 和 P.Chen，《Voice2series：为时间序列分类重新编程声学模型》，载于*ICML*。2021 年，第 139 卷，机器学习研究论文集，第 11808–11819 页，PMLR。[9]H.Yen、P.Ku、C. H.Yang、H.Hu、S. M.Siniscalchi、P.Chen 和 Y.Tsao，《基于对抗式重新编程的低资源语音命令识别研究》，*CoRR*，第 abs/2110.03894 卷，2021 年。

[10] R.Vygon 和 N.Mikhaylovskiy，“利用三元组损失学习用于关键词检索的高效表示”，载于国际语音与计算机会议，2021 年，第 773–785 页。

[11]Y.Tian 和 P. J.Gorinski，“利用 Reptile 改进端到端语音意图分类”，载于*INTER-SPEECH*。2020 年，第 891–895 页，ISCA。

[12] D.Bermuth、A.Poeppel 和 W.Reif，“Finstreder：使用现代语音转文本模型，通过有限状态转换器实现简单快速的口语理解”，*arXiv* 预印本 *arXiv:2206.14589*，2022 年。

[13] J.Shor、A.Jansen、W.Han、D.Park 和 Y.Zhang，“使用自监督 conformer 的通用副语言语音表示”，载于*ICASSP*，2022 年，第 3169–3173 页。

[14] A.Ahamad、A.Anand 和 P.Bhargava，“AccentDB：用于辅助神经语音识别的非英语母语口音数据库”，载于*LREC*。2020 年，第 5351–5358 页，欧洲语言资源协会。

[15] S.Castro、D.Hazarika、V.P´erez-Rosas、R.Zimmermann、R.Mihalcea 和 S.Poria，“迈向多模态讽刺检测（显然是一篇完美的论文）”，载于*ACL* 第 *57* 届会议论文集，2019 年，第 4619–4629 页。

[16] A.Ray、S.Mishra、A.Nunna 和 P.Bhattacharyya，“用于识别讽刺中情绪的多模态语料库”，载于第十三届 *LREC* 论文集，2022 年，第 6992–7003 页。

[17] K.Hechmi、T. N.Trong、V.Hautam¨aki 和 T.Kinnunen，“用于年龄和性别识别的 VoxCeleb 扩充”，载于*ASRU*，2021 年，第 687–693 页。

[18] F.Jia、S.Majumdar 和 B.Ginsburg，“MarbleNet：用于语音活动检测的深度 1D 时间通道可分离卷积神经网络”，载于 *ICASSP*，2021 年，第 6818–6822 页。

[19] K.Chen、X.Du、B.Zhu、Z.Ma、T.Berg-Kirkpatrick 和 S.Dubnov，《 Hts-at：一种用于声音分类和检测的分层令牌语义音频变换器》，载于*ICASSP*，2022 年，第 646–650 页。

[20]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。

[21] V.Renkens、S.Janssens、B.Ons、J. F.Gemmeke 等，“使用弱监督非负矩阵分解获取序数词”，载于*SLT*，2014年，第30–35页。

ˇ[22] A.Kolesau 和 D.Seˇsok，“用于语音激活的无监督预训练”，《style id='2'>《应用科学》，第10卷，第23期，第8643页，2020年。

[23] L. T.Benamer 和 O. A.Alkishriwo，“阿拉伯语语音命令识别数据库”，载于*CEST*，2020年。

[24]L.Lugosch、M.Ravanelli、P.Ignoto、V. S.Tomar 和 Y.Bengio，“用于端到端口语理解的语音模型预训练”，载于*INTERSPEECH*。2019年，第814–818页，ISCA。

[25] K.MacLean，《Voxforge》，2018，*KenMacLean.[在线]*。网址：*http://www.voxforge.org/home*。*[访问日期：2012]*。

[26]A.Nautsch、X.Wang、N.Evans、T. H.Kinnunen、V.Vestman、M.Todisco、H.Delgado、M.Sahidullah、J.Yamagishi 和 K. A.Lee，《ASVspoof 2019：用于检测合成语音、转换语音和重放语音的欺骗攻击对策》，*IEEE* 生物识别、行为与身份科学汇刊，第 3 卷，第 2 期，第 252–265 页，2021 年。

[27] C.Busso、M.Bulut、C.-C.Lee、A.Kazemzadeh、E.Mower、S.Kim、J. N.Chang、S.Lee 和 S. S.Narayanan，《Iemo-cap：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 4 期，第 335–359 页，2008 年。

[28] A.Nagrani、J. S.Chung 和 A.Zisserman，“VoxCeleb：大型说话人识别数据集”，载于*INTER-SPEECH*，2017 年。[29]E.Fonseca、J.Pons、X.Favory、F.Font、D.Bogdanov、A.Ferraro、S.Oramas、A.Porter 和 X.Serra，“Freesound 数据集：创建开放音频数据集的平台”，载于*ISMIR*，2017 年，第 486–493 页。[30] K. J.Piczak，“ESC：用于环境声音分类的数据集”，载于第 *23* 届 *ACM* 国际多媒体会议论文集，2015 年，第 1015–1018 页。[31]W.Hsu、B.Bolte、Y. H.Tsai、K.Lakhotia、R.Salakhutdinov 和 A.Mohamed，“HuBERT：通过隐藏单元的掩码预测进行自监督语音表征学习”，发表于*IEEE ACM* 音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021 年。[32] C.Liu、H.Lee 和 W.Yih，“结构化提示调优”，*CoRR*，第 abs/2205.12309 卷，2022 年。

[33] S. M.Lundberg 和 S.-I.Lee，“一种统一的模型预测解释方法”，神经信息处理系统进展，第 30 卷，2017 年。