arXiv:2305.03058v1 [eess.AS] 3 May 2023

# 即插即用多语言少样本口语词识别

AAQIB SAEED，埃因霍温理工大学，荷兰VASILEIOS TSOUVALAS，埃因霍
温理工大学，荷兰

随着技术进步和数字设备日益普及，无缝的人机通信正变得越来越重要。移动设备、可穿戴设备及其他物联网（IoT）设备的广泛应用，改变了我们与这些智能设备的交互方式，使准确的口语词识别成为实现有效交互的关键。然而，构建能够处理新颖关键词的稳健口语词检测系统仍然具有挑战性，尤其是在训练数据有限的低资源语言中。本文提出 PLiX，这是一种多语言、即插即用的关键词检出系统，利用小样本学习挖掘海量真实世界数据，并在测试时识别未见过的口语词。我们的少样本深度模型通过 20 种语言中数百万个一秒长的音频片段进行训练，在实现最先进性能的同时保持了极高的效率。大量评估结果表明，只需提供一个支持样例，PLiX 就能泛化到新颖的口语词，并且开箱即用地适用于未见过的语言。我们发布了模型和推理代码，以支持未来的研究，以及面向新兴设备的语音用户界面开发。

CCS 概念：• **计算方法** → **学习范式**；• **以人为中心的计算** → 普适与移动计算理论、概念与范式。

Addi关键词与短语：小样本学习、关键词检出、语音、音频、深度神经网络 ks

## 1 引言

移动设备、可穿戴设备及其他物联网设备的普及，改变了我们的沟通、导航和执行各种任务的方式，使其成为日常生活中不可或缺的一部分。日常生活中采用这些新兴设备，要求人与机器之间能够无缝沟通。因此，准确识别口语词[1], 或关键词检出（KWS），已成为人与此类设备交互的基础，因为它能够促进自然对话。具体而言，它是 Siri 和 Alexa 等虚拟助手的重要组成部分。这些助手彻底改变了用户与设备交互的方式，使体验更加自然、个性化且直观。关键词检出器是资源需求相对较低的唤醒词引擎，会持续监控音频，监听音频流中的特定触发短语。一旦检测到该短语，系统便会唤醒并启动服务或执行任务，例如打开电视，或在驾驶时与车载系统进行交互。

随着深度学习算法和语音识别技术的发展，关键词检出已得到广泛应用，使用户与设备之间的交互更加顺畅。这项技术正越来越多地应用于各种领域，包括智能家居、汽车系统和医疗保健。然而，为多种语言构建稳健且准确的模型仍是一项艰巨的任务，尤其是对于缺乏大量训练用标注数据的低资源语言而言。尽管基于深度神经网络的口语词识别取得了重大进展，传统学习方法仍需要大量训练数据，这既耗时又昂贵，而且为适应不断变化的用户需求，还需要重新训练模型以纳入新的关键词。同样，传统学习方法可能需要大量计算资源，因此不适用于资源受限设备，也无法实现设备端个性化。因此，有必要开发更先进、更有效的关键词检出系统，以适应多语言世界，并能在推理时快速调整以识别新词。

[1]我们交替使用单词、术语和类别。

作者地址： AAQIB SAEED，a.saeed@tue.nl，埃因霍温理工大学，埃因霍温，荷兰；Vasileios Tsouval as，埃因霍温理工大学 ，埃因霍温，荷兰。

开发此类口语词识别系统对于塑造人机交互的未来可以发挥至关重要的作用，使不同语言背景下的人们与数字设备之间能够实现顺畅、自然的交流。

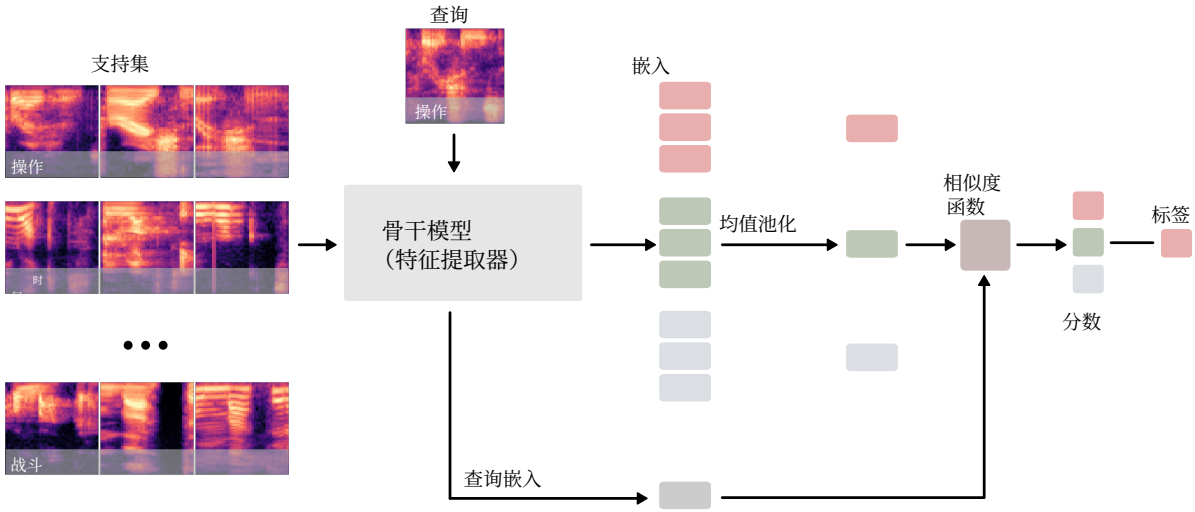



图1. 原型网络的小样本训练范式示意图 [21]，用于多语言口语词识别。

近期研究提出了能够学习深度模型的方法，使其能够在训练数据有限的情况下泛化到未见过的任务。这种方法称为小样本学习。在这一领域，分类器必须仅凭每个类别的少量样本，学会识别新类别。传统的小样本学习技术将问题设定为封闭集分类任务，目标是将查询样本标记为某个 $C$已知类别 $K$，每个类别仅提供若干样本。这种方法在解决从声音事件检测到 [20, 27] 图像分类 [14, 17, 21, 22]等元学习问题方面已取得巨大成功。其核心思想是让模型通过少量示例进行学习，从而快速适应新任务，使该方法能够应对不同领域中的各种任务。为解决缺乏通用、语言无关且易于使用的关键词检测器这一问题，我们提出了**PLiX** (示意图见图1 )，这是一个利用小样本学习挖掘海量真实世界数据、实现未见过的口语词识别的系统。总体而言，我们的工作有以下主要贡献：

- 我们开发了 PLiX，这是一种通用型、多语言且即插即用的少样本关键词检出系统，使用超过 1200 万个以 16 千赫采样的一秒音频片段进行训练和评估。
- 利用最先进的神经网络架构，学习性能出色且可学习参数更少、效率更高的少样本模型。
- 我们开展了范围广泛的评估，系统地量化该系统在 20 种语言和数千个类别（即单词或词语）上的效果；结果表明，在测试时，每个类别仅需一个支持样例，系统便能泛化至未见过的单词。
- 我们证明，我们的模型在 5 种未见过的语言上进行单样本学习时，具有出色的泛化能力。此外，在具有挑战性的 FLEURS 基准上进行跨任务迁移评估时， [7], 我们的模型无需重新训练，便能出色地完成语言识别。
- 为了作为未来利用元学习开展口语词汇检测研究的基础模块，并推动产品开发，我们在以下地址发布模型权重和推理代码：https://github.com/FewshotML/plix。

# 2 方法论

## 2.1 预备知识

*2.1.1* 小样本学习。小样本学习（FSL）是机器学习的一个子领域，研究如何在仅有少量标注数据的情况下学习执行任务（例如目标分类）。在经典学习范式中，主要目标是利用大量标注训练数据，获得能够对同类别中此前未见过的数据进行良好泛化的模型。然而，在许多现实世界应用中，获取大量标注数据往往既艰巨又昂贵、耗时，因此 FSL 成为开发数据高效模型的一种颇具吸引力的范式。少样本学习器的优势在于能够凭借极少的标注数据快速适应新任务，从而为标准学习方法提供了一种实用且有吸引力的替代方案。近年来，FSL 通过采用情节式范式取得了进展。具体而言，设想我们拥有一个针对某一特定类别集合 $C_{train}$的大型标注数据集。然而，我们的最终目标是为一组不同的新类别 $C_{test}$创建分类器，而这些类别仅有数量有限的标注样例。对此，情节式范式提供了一个框架，用于模拟测试期间可能遇到的各类小样本问题，同时利用类别 $C_{train}$所拥有的丰富标注数据。

在少样本学习范式中，模型的训练是在 $K$-shot、 $N$-way 的任务回合上进行的。首先，从可用类别集合中抽取 N 个类别组成一个小子集 $C_{train}$。随后，生成一个训练（支持）集 $S$ 和一个测试（查询）集 $Q$。支持集 $S$ 包含来自 N 个类别中每个类别的 K 个样本，即 $S = \{(x_1, y_1) \ldots, (x_{N\times K}, y_{N\times K})\}$，其中 $x_i \in \mathbb{R}^D$表示维度为 D 的输入向量，$y_i \in 1$、...、 $N$表示类别标签。另一方面，测试集 $Q$由来自相同 N 个类别的不同样本组成，即 $Q = \{(x_1^*, y_1^*) \ldots, (x_T^*, y_T^*)\}$，其中 $x_i^* \in \mathbb{R}^D$、 $y_i^* \in \{1, 2, \ldots, N\}$，且 $T$表示查询样本的数量。为了在此类任务回合上训练，模型 $f_\theta(\cdot)$ 接收支持集 $S$，并更新其参数，以最小化指定的损失函数 $\min_\theta \sum_{E\in D} L(f_\theta(S)\ Q)$。该损失函数衡量模型对查询集 $Q$中样本的预测与真实标签之间的差异。这一训练过程通过迭代调整模型的内部表示，提升模型在测试时仅凭有限标注数据对新类别进行泛化的能力。因此，这一框架下的学习通常称为学会学习或元学习 [24], ，因为它旨在训练模型，使其能够凭借有限的标注数据快速适应未见过的任务，这类似于人类利用已有知识和经验，通过少量示范学习新任务的方式。

*2.1.2* 原型网络。原型网络（或 ProtoNet）是一种简单但非常有效的小样本学习方法，由 Snell 等人提出 [21]。原型网络的核心概念是为支持集中每个类别计算一个原型。这些原型作为类别的代表性概括，随后用于对查询集中的实例进行分类。从本质上说，原型网络将输入样本映射到度量（或嵌入）空间，并利用原型与查询示例之间的距离进行分类。

学习过程首先计算支持集 $S$中每个类别的原型。令 $S_i$表示 $S$ 中属于类别 $i$的实例集合，令 $f_\theta$ 表示一个具有可学习参数$\theta$的神经网络。类别 $i$的原型定义为其在 $S_i$中的实例的平均嵌入：$\mathrm{p}i = \frac{1}{|S_i|} \sum(x,y)\ \epsilon S_i\ f_\theta(x)$，其中 $|\cdot|$ 表示集合的基数。给定原型 $\mathrm{p}_i$，查询集中的实例 x 的分类得分可计算为$\overline{\mathrm{x}}$ 与各原型 $\mathrm{p}_i$之间的负欧几里得距离：$\mathrm{s}_i = -\|f_\theta(\mathrm{x}) - \mathrm{p}_i\|_2^2$，其中 $\|\|_2$ 表示欧几里得范数。最后，x 的预测类别是分类得分最高的类别：$\hat{}y = \mathrm{argmax}i\epsilon\{1,...,N\}\ s_i$。ProtoNet 可以通过最小化查询集中各实例预测类别概率的交叉熵损失并执行梯度

下降更新来进行端到端训练。令 $Q_i$表示查询集中属于类别 $i$的实例集合，令 $y_j \in 1、\ldots\ldots、 N$表示实例 $\mathbf{x}j$ 在 $Q$ 中的真实类别标签，其中$\mathbf{1}(\cdot)$为示性函数。于是，单个任务回合的损失可写为：$\mathrm{L} = -\sum_{j=1}^{|Q|}\sum_{i=1}^{N} 1(y_j = i) \log \frac{\exp(s_i)}{\sum_{k=1}^{N}\exp(s_k)}$。

原型网络通过利用根据支持集样例计算得到的一组原型，学习对查询集中的实例进行分类。训练集完成后，其泛化性能将在测试集任务回合上进行评估，这些任务回合包含来自 $C_{test}$而非 $C_{train}$中类别的实例。对于每个测试集任务回合，原型网络使用支持集 $S$生成的预测器，将每个查询输入$\mathbf{x}^*$分类为最可能的类别$\hat{y} = \mathrm{argmax}_c\, p(c|\mathbf{x}^*, \mathbf{p}_i)$。该方法在小样本学习基准测试上展现了最先进的性能。由于其简单性和可扩展性，我们利用原型网络来解决真实环境中的口语词识别任务。

## 2.2 少样本识别系统

*2.2.1* 概述。 图 2 展示了 PLiX 系统设计的高层概述。该系统由一个声学模型组成，这是一种利用声音的视觉表示，从海量多语言音频数据中学习有用特征的神经网络（第 2.2.2 节）。我们基于原型网络构建的小样本学习系统，通过对口语词进行情景式训练，学习如何泛化到测试时未见过的词语（第 2.2.3 节）。该系统使用以 16 kHz 采样的一秒音频片段作为查询和支持样例，以检测感兴趣的口语词。所设计的系统可应用于各种场景（第 2.2.4 节），以即插即用的方式开发以语音为界面的新型应用。

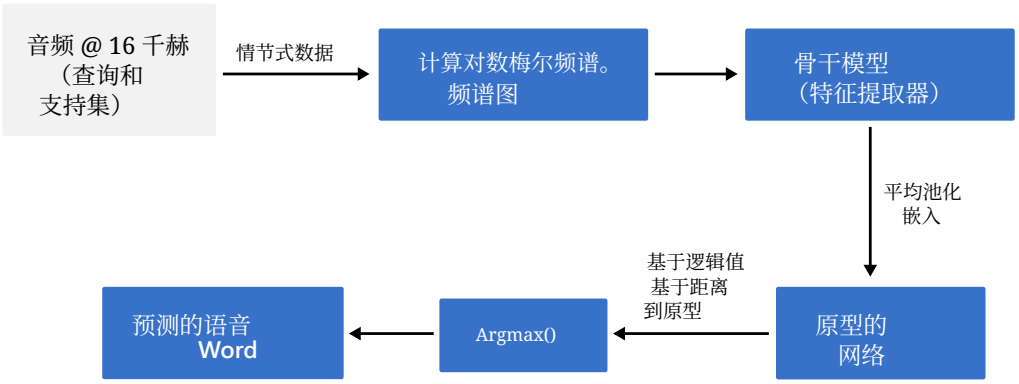



图 2. PLiX 系统设计概览。

*2.2.2* 声学模型架构。 我们利用最先进的神经网络架构作为“编码器”或“特征提取器”，从音频中学习具有泛化能力的表征。我们重点关注在快速训练和推理方面均表现出色的神经网络架构（包括在资源受限设备和云端上的应用）。我们的少样本学习器使用 EfficientNet-v2 [23] 和 TinyNet [9] 作为原型网络中的编码器或骨干网络。根据模型规模，我们进一步将前者称为“基础”模型，将后者称为“小型”网络。对于模型输入，我们使用音频信号的视觉表示，即“对数梅尔频谱图”。我们通过计算音频信号的短时傅里叶变换，再使用一组梅尔频率滤波器对所得频谱进行滤波来获得该表示。随后计算所得梅尔滤波器组能量的对数，从而得到一种紧凑且信息丰富的表示，用于描述音频信号随时间变化的频谱内容。与其他频谱表示相比，对数梅尔频谱图具有多项优势，包括能够捕捉信号的频谱特征和时域特征，以及对噪声和失真的鲁棒性。具体而言，它已

被广泛应用于语音识别、音乐流派分类和声音事件检测等多个领域。

EfficientNet-v2 旨在以比前代更少的参数和更快的训练速度，实现更优的性能。它引入了多项关键改进，包括一种称为“复合缩放”的缩放方法，以较小、较快的模型实现更优性能。复合缩放结合了深度、宽度和输入分辨率的缩放，从而在准确率和计算效率之间找到最佳平衡。此外，它还引入了一种名为“Fused-MBConv”的新型卷积块，将多种卷积层合并为一个块。与 EfficientNet 中使用的原始 MBConv 块相比，Fused-MBConv 的效率更高，因为它消除了冗余操作并减少了参数数量。主架构有多个不同变体，主要区别在于参数数量和所用的预训练数据集。由于泛化能力更强，我们使用包含 52.8 百万个参数的 EfficientNet-v2（Medium）作为“基础”模型，用于训练多语言和仅英语的少样本模型。

为了设计体积更小、性能却依然出色的少样本学习器，我们采用 TinyNet 作为“轻量级”模型，利用其极小的内存占用来训练特定口语语言的模型。TinyNet 是一系列神经网络，旨在以较少的参数实现较高的准确率。这些网络尤其适合部署在资源受限设备上，例如智能手机或嵌入式系统，因为这类设备的内存和处理能力有限。TinyNet 的主要创新在于对网络的三个关键维度进行调控：分辨率、深度和宽度。该方法称为“模型魔方” [9], ，其中每个维度对应魔方的一个面，旋转这些面即可生成大小和准确率各异的网络配置。在我们的工作中，我们采用最小的模型 TinyNet-E，包含 761,396 个参数，将其作为原型网络使用的骨干网络“特征提取器”。

*2.2.3* 训练流程。如前所述，我们使用重采样至16kHz 的音频，其中每个实例均包含一个持续一秒的口语词。我们从音频片段中提取对数梅尔频谱图（可视化结果见图3 ），并将其作为模型的一部分在 GPU 上执行，以便高效处理大规模音频数据集。具体而言，我们使用长度为 400 的窗口，步长为 160，并在 64 个梅尔频率刻度频点上提取特征，频率范围为 60-7800Hz。对于所考虑的两种编码器网络，我们均在应用全局平均池化后，从倒数第二层提取特征，最终得到维度为 1280 的特征向量。我们利用这些特征计算平方欧几里得距离，作为查询集与支持集中各实例之间的相似度度量，以实现分类。为了生成训练任务回合，我们随机采样类别（即口语词），然后采样与这些类别对应的音频片段，以构建查询集和支持集。在多语言模型训练的情况下，我们将特定语言的数据集合并，形成一个更大的数据池，并在训练阶段采用前文所述的策略生成任务回合。这种方法大幅简化了整个训练流程，并生成包含不同语言数据的多样化任务回合，可将其视为一种隐式的多任务学习方式 [2]。我们采用 Adam 优化器，以 0.001 的学习率训练模型，共进行 100K 个任务回合，但对于「基础」多语言模型，由于其使用了规模更大的合并训练集，因此训练 200K 个任务回合。除非另有说明，所有情况下训练期间的查询数量均设置为10。此外，我们使用 ImageNet 预训练权重进行初始化，因为我们发现，与随机初始化相比，这种方式能够实现更快的收敛。我们认为，使用基于音频数据集预训练的模型可以进一步改进初始化效果，这将作为未来的研究方向。

*2.2.4* 推理策略。 如第 2.2.3 节所述，一旦通过口语词的情节式数据训练出少样本模型，便可通过多种方式利用该模型，在真实环境中检测音频关键词，并解决其他相关问题（例如口语语言检测）。在测试时，用户提供查询样例（即音频片段），这些样例可能包含待检测的目标关键词，同时还提供少量支持样例。如果音频片段长度超过一秒，可以将其切分为一秒长的片段，并将每个生成的片段视为一个查询样例。随后，少样本模型可以计算查询样例与支持集中的样例之间的相似度，并根据原型距离分配标签。值得注意的是，原型网络不要求具有相似的配置

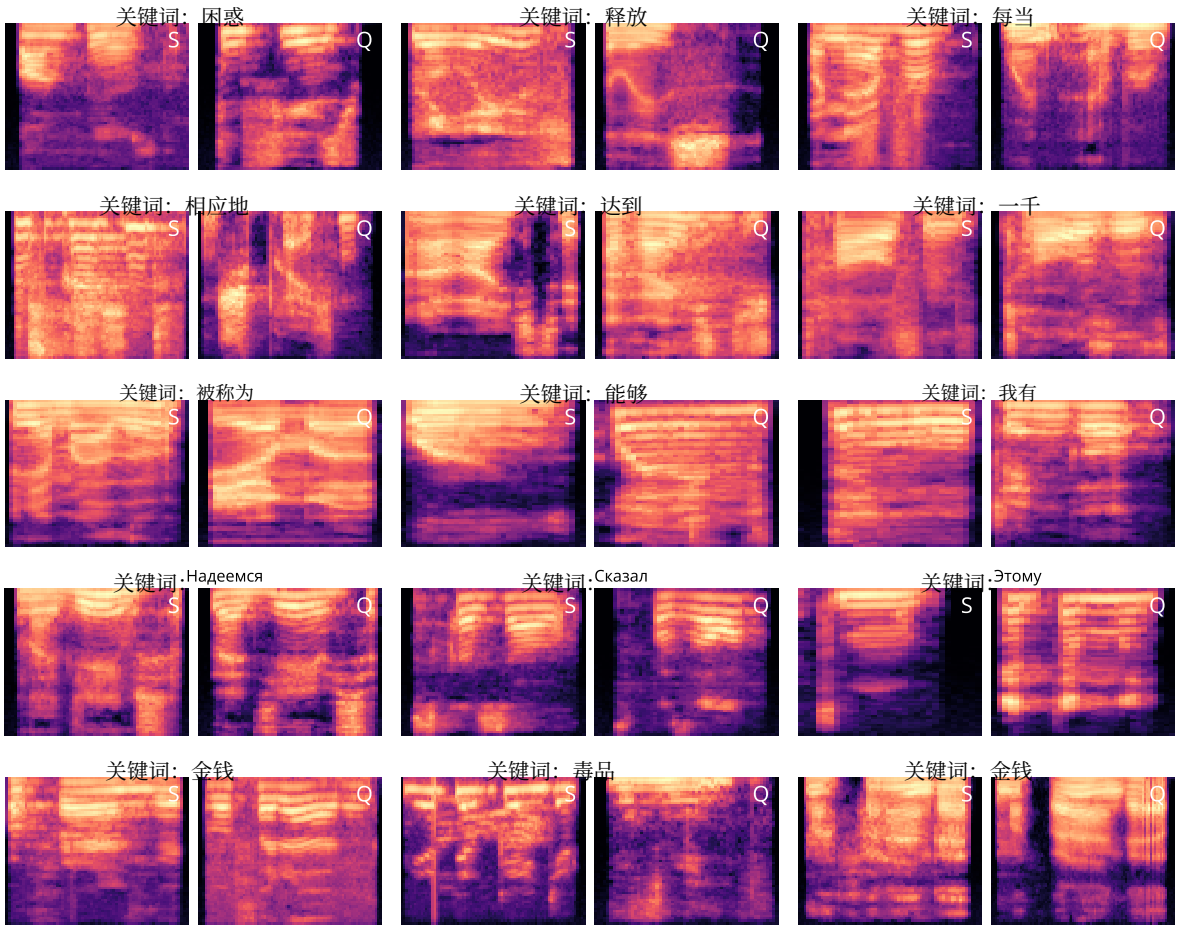



图 3。我们展示了一个任务回合中随机选取的查询 $Q$ 和支持 $S$ 集示例，该任务回合包含来自五种语言（即“en”、“de”、“es”、“ru”和“rw”）的不同关键词。对一秒钟的音频片段计算对数梅尔频谱图，使用 64 个梅尔频率刻度频点。

推理时查询和支持集的数量可像训练时那样灵活调整。例如，在训练时，我们可以学习一个 $N$路 $K$样本的模型，其中 可以是 $N = 10$ 和 $K = 5$ 但在推理时，我们可以同时改变 $N$和$K$。然而，请注意，如果 $N$ 和 $K$ 与训练配置相比变化都很大，性能可能会显著下降。此外，在单样本设置中，用户在测试时只提供一个支持样例，我们可以利用推理时数据增强 [27]来生成多个支持样例。

## 2.3 即插即用方法和开源

我们的方法旨在简化口语词识别模型在现实场景中的部署。通常，开发和维护能够可靠识别口语关键词的标准监督模型颇具挑战，因为用户对所关注口语词的需求可能会随时间变化。因此，我们的少样本模型提供了一种可行的方式，只需少量标注数据（以支持样例的形式），便能持续支持对新口语词的检测

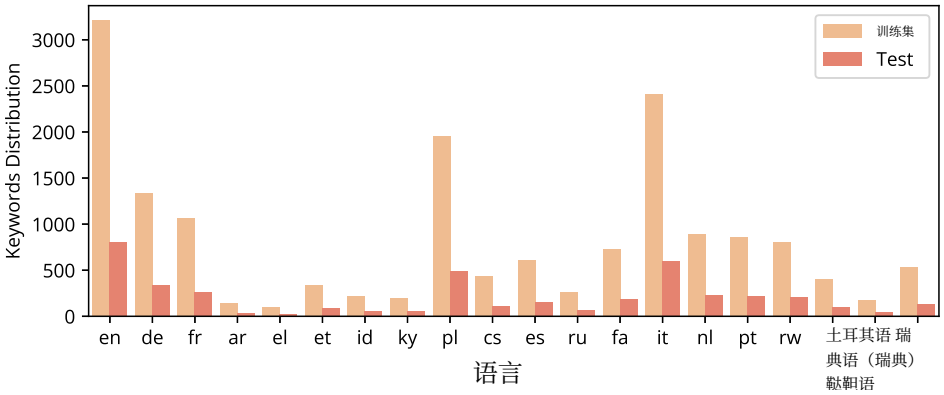



图 4. 所考察语言中“训练集”和“测试集”里的唯一关键词（类别或口语词） es.

。此外，我们的即插即用设计便于集成到现有系统中，只需进行极少修改；而且只需少量支持集样例，即可支持多种语言的识别。为了让其他研究人员能够探索这一领域，并让开发者可以轻松地将其用于未来的应用，我们还发布了推理代码和模型权重，包括多语言模型和特定语言模型。通过此次发布，我们还旨在促进针对新用例的适配与扩展。我们的预训练模型可用于低资源语言的迁移学习。它们能够实现较高的准确率，同时免去在大型数据集上进行大量训练的需要，而这一过程既耗时又耗费资源。

## 3 实验

### 3.1 多语言口语词数据集

我们使用多语言口语词语料库（MSWC） [12]来训练和评估我们的小样本学习模型。该语料库旨在用于学术研究和商业应用，尤其是关键词检出领域。MSWC 是一个庞大且不断扩展的音频录音集合，其中包含以50种不同语言录制的口语词。这些语言的使用者总数超过5 十亿。具体而言，其中包含超过340万个关键词对应的一秒口语词音频片段，共计23.4 百万个样本。该语料库通过对众包语音语料库 Common Voice 数据集进行强制对齐自动生成 [1]。由于关键词数量及其对应音频片段众多，它非常适合进行情景式训练，因为每个任务回合都隐含要求包含新类别。为便于基准测试，该语料库默认划分为训练集、开发集和测试集，比例分别为80%、10%和10%。数据集中的音频片段采用单声道、48KHz 采样率进行 Opus 压缩。在我们的工作中，我们在预处理阶段将所有音频文件转换为 WAV 格式，并重采样至16KHz。我们聚焦于 MSWC 语料库中的20 种高资源和中资源语言。对于高资源语言“en”“de”“es”“fr”“fa”“ru”和“rw”，我们过滤掉音频片段少于200 个的关键词。同样，对于其余纳入考虑的语言，如果音频片段数量少于25个，我们也会移除相应关键词。需要注意的是，少样本模型使用模型在训练阶段未接触过的新类别（或不相交类别）集合进行评估。为此，我们首先合并语料库提供的“训练集”和“开发集”关键词，然后按照80 : 20 的比例基于关键词进行划分，以确保训练集和测试集中的关键词不重叠。图4展示了各语言中唯一关键词的分布，其中英语的关键词数量最多。此外，图5展示了我们生成的数据划分中的音频片段（每段一秒）

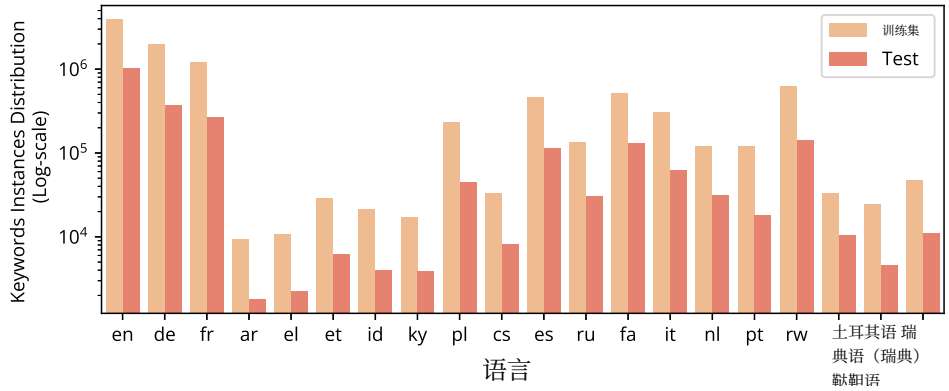



图5 数量 每种语言的音频片段数量，突出了 MSW 所体现的口语语言资源丰富程度 C.

（每段一秒）5。我们为小样本学习和评估生成的数据集总共包含超过12 百万个一秒音频片段，其中训练集和测试集分别包含16、665 和4、174 个唯一关键词。

## 3.2 评估协议

我们在测试集中未见过的口语词（或类别）上评估少样本模型。我们使用 1000 个任务回合进行测试，除非另有说明，每个任务回合包含 15 个查询。需要注意的是，我们会系统地改变类别数量和支持样例数量，并在后续章节讨论相应结果时说明这些设置。测试任务回合的构建方式与训练阶段保持一致。具体而言，我们随机抽取类别标签，并针对每个类别抽取指定数量的支持实例和查询实例。在测试过程中，我们通过基于与原型之间的最近距离为查询样例分配标签，来识别测试录音中的口语关键词。我们报告所有实验条件下全部任务回合的平均准确率。

## 3.3 结果

单一多语言模型的能力：我们首先考察多语言“base”模型（即 EfficientNet-v2）在一系列少样本评估场景中的泛化能力。具体而言，我们系统地改变类别数（$N$）以及每个类别的支持样例数（$K$），以确定模型在所考察语言上对未见过的词语的泛化表现。在实验中，我们首先仅使用各语言训练集中的词语，将基础模型训练为 10 类 $\times 5$ 样本分类模型，然后在留出的词语上评估训练好的模型。结果如图 6 所示。对于每类只有一个支持样例的 5 类、10 类和 15 类分类，我们发现模型表现出色，准确率超过 80%；在大多数情况下，5 类分类的准确率超过 90%。即使在难度更高的 25 类分类中，模型在英语和德语等资源丰富的语言上也能实现准确识别。此外，对于 10 类分类，将支持样例数增加到 2 个后，检测率始终显著提高。然而，对于卢旺达语（rw）和瑞典语（瑞典）（sv-SE）等语言，25 类 $\times 1$ 样本分类的准确率相对较低，分别为 66% 和 71%。基于此，我们将支持样例数增加到 5 个，这使“rw”的准确率大幅提升至 87.6%，而“sv-SE”的准确率提升至 88.9%，展现了即使仅增加少量样例，也能通过扩大支持集规模改善具有不同声学和结构特征的语言的表现。总体而言，这些结果凸显了我们的多语言基础模型强大的泛化能力，尤其是考虑到

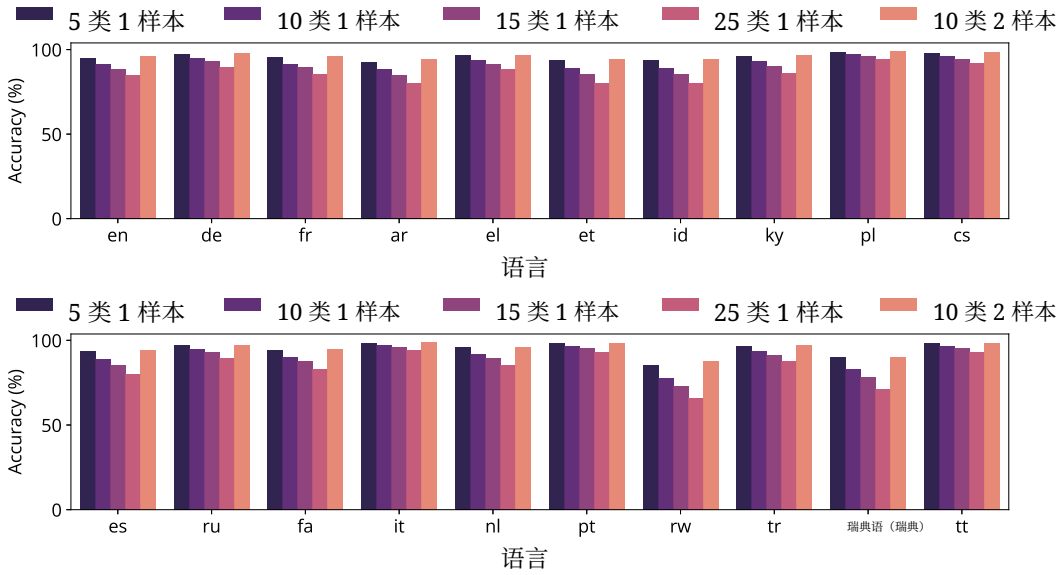



图 6. 对在所考察的 20 种语言数据上训练的多语言“基础”模型进行性能评估。少样本模型针对 10 路（$N$）和 5 样本（$K$）分类任务进行训练。在推理时，我们改变 $N$ 和 $K$，以展示即使只提供支持样例时的泛化能力。

数据有限。该模型的性能表明，使用单一模型进行口语词少样本识别具有良好的应用潜力，尤其是随着支持样例数量的增加，有望缓解低资源语言所面临的挑战。

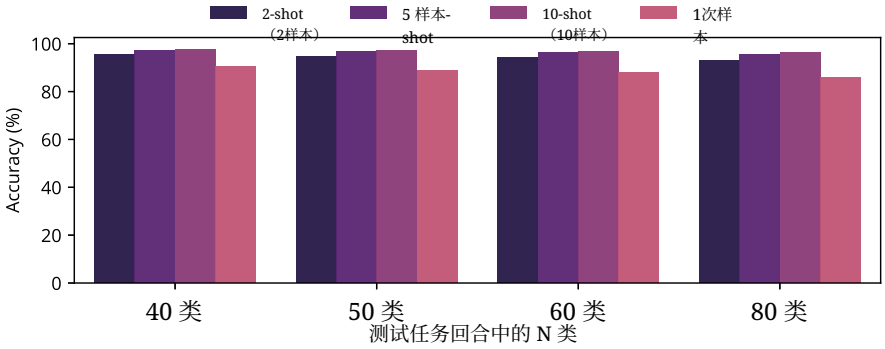



图 7. 使用 30 路、5 样本任务训练的少样本英语专用“基础”模型的结果。我们旨在强调，在测试时，同一模型可以处理大量分类任务，且支持集样例数量各不相同。

**类别数增加时的准确分类**：随后，我们对仅使用英语口语词训练的“基础”模型进行了实证评估。我们在一个 30 类 × 5-shot 任务上训练少样本模型，并在测试期间针对不同的 $N$ 和 $K$ 设置对其进行评估。具体而言，我们将 $N$ 类（类别数量）

分类任务的类别数从 40 增至 50，再增至 60，最终增至80 类。这表明，模型训练完成后，即使类别数量增加而支持样例数量较少，它仍能从一组其他词语中区分出某个口语词。我们在图7 中给出了这些结果。首先，当$N$ 设为 40 ，且每个类别只有一个支持样例时，模型的准确率达到 90% ；而在 80% 的 80类分类任务中，准确率仅下降四个百分点。将支持样例数量从一个增加到十个后，我们观察到，在所有考察的实验条件下，检测率均逐步提高。当支持样例为五个时，在更具挑战性的 60 类和 80 类任务中，准确率分别达到 96% 和 95%。

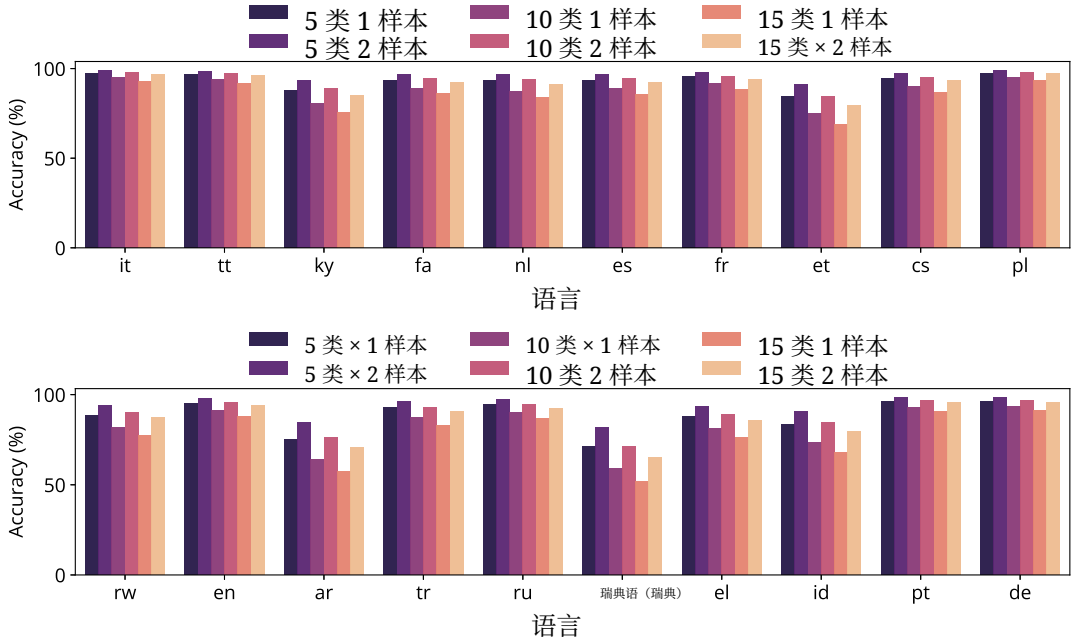



图 8. La 不同查询数量和支持集样例数量下的特定语言“小型”模型评估。

小巧而强大的语言专属少样本模型：在图8中，我们展示了针对每种所考虑的语言分别独立训练和评估的“小型”语言专属模型（即 TinyNet-E）的结果。我们开发这些少样本模型时，尽可能缩小其内存占用和计算资源需求，因为具有泛化能力的模型适用于在边缘设备和可穿戴设备上进行推理。这些模型在一个10类 × 5样本识别任务上进行训练。我们在不同类别数和样本数的设置下对它们进行评估。我们发现，对于资源丰富的语言，“较小型”模型的泛化效果出色；例如，对于意大利语（‘it’）、英语（‘en’）、德语（‘de’）和西班牙语（‘es’），在每类仅有一个支持样例的5类任务中，准确率高于95%。即使将分类任务扩展到区分15 个口语词，且每类有 2 个支持样例时，大多数情况下的表现仍高于80% 。然而，对于数据集规模相对较小的其他语言，模型表现较差。当支持样例较少、类别数较多时，阿拉伯语（‘ar’）、瑞典语（‘sv-SE’）、印度尼西亚语（‘id’）和爱沙尼亚语（‘et’）的表现尤为明显。请注意，我们在数据极度稀缺的设置下评估模型，即仅提供一个或两个支持集样例。不过，当我们将印度尼西亚语（‘id’）10类任务的支持样例数增加到5个时，准确率达到

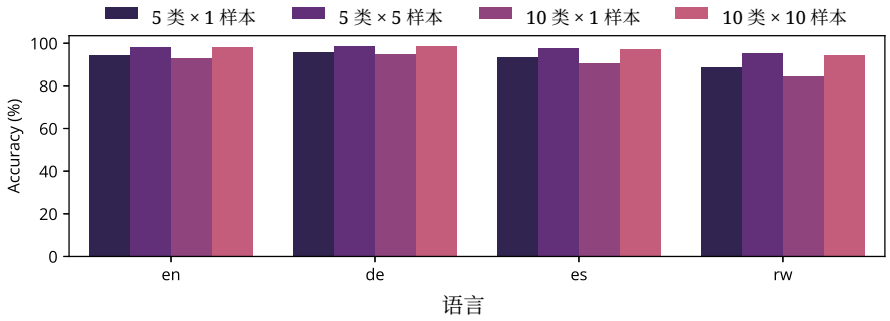



图 9. 在训练和测试设置中，‘small’模型作为少样本学习器，在四种资源丰富的语言上针对不同的 $N$ 和 $K$ 所取得的实验结果；也就是说，针对 5 类 × 1-shot 任务训练的模型，在推理时以相同配置进行评估。

90%。我们在阿拉伯语（‘ar’）中也观察到了同样的效果：仅将支持样例数从一个增加到五个，准确率就从 63% 提升到 85%。根据这些结果，我们得出结论：只要训练用的情节式数据足够丰富，或者在推理时提供更多支持样例，较小型模型也能很好地识别未见过的口语词。

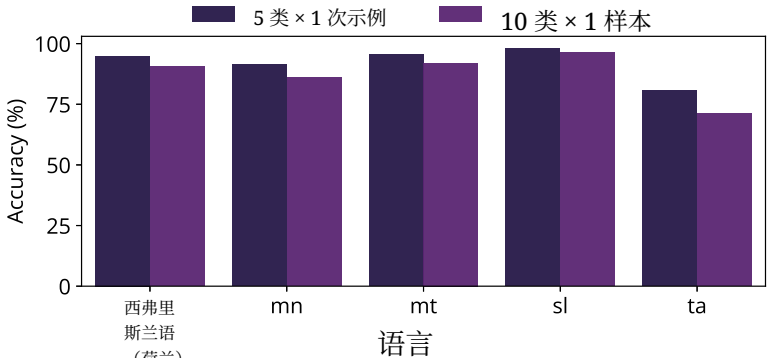



图 10. MSWC 中 5 种未见过的低资源语言上的实验结果，分别为弗里西语、蒙古语、马耳他语、斯洛文尼亚语和泰米尔语。我们采用单样本方式评估“base”多语言模型，以展示我们所提出的少样本模型的泛化能力。

少样本参数变化：为了在 MSWC 背景下进一步评估我们使用资源丰富的语言训练的少样本模型的泛化能力，我们在相同配置下训练和测试“small”模型，具体而言，保留相同的 $N$路和 $K$样本参数。如图9所示，我们观察到模型性能显著提升，尤其是“rw”，其准确率达到84% ，在10路任务中仅使用一个支持样例便取得了这一成绩，较图8所示评估设置和结果提高了三个百分点。尽管其他语言的性能提升较为有限，但这表明了一个值得关注的现象：即使在测试阶段改变 $N$ 和 $K$ ，少样本模型仍能以较高准确率识别口语词；从实际部署的角度来看，这是一个宝贵的特性。

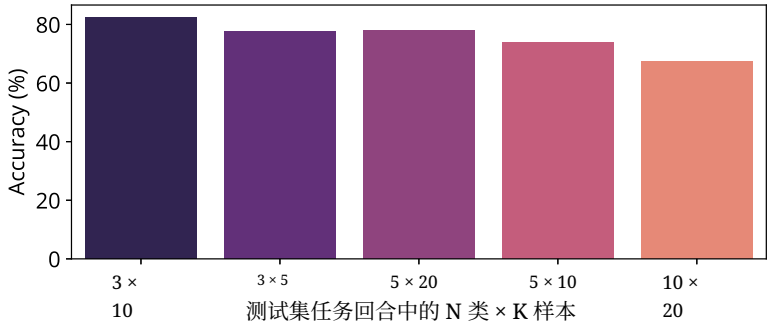
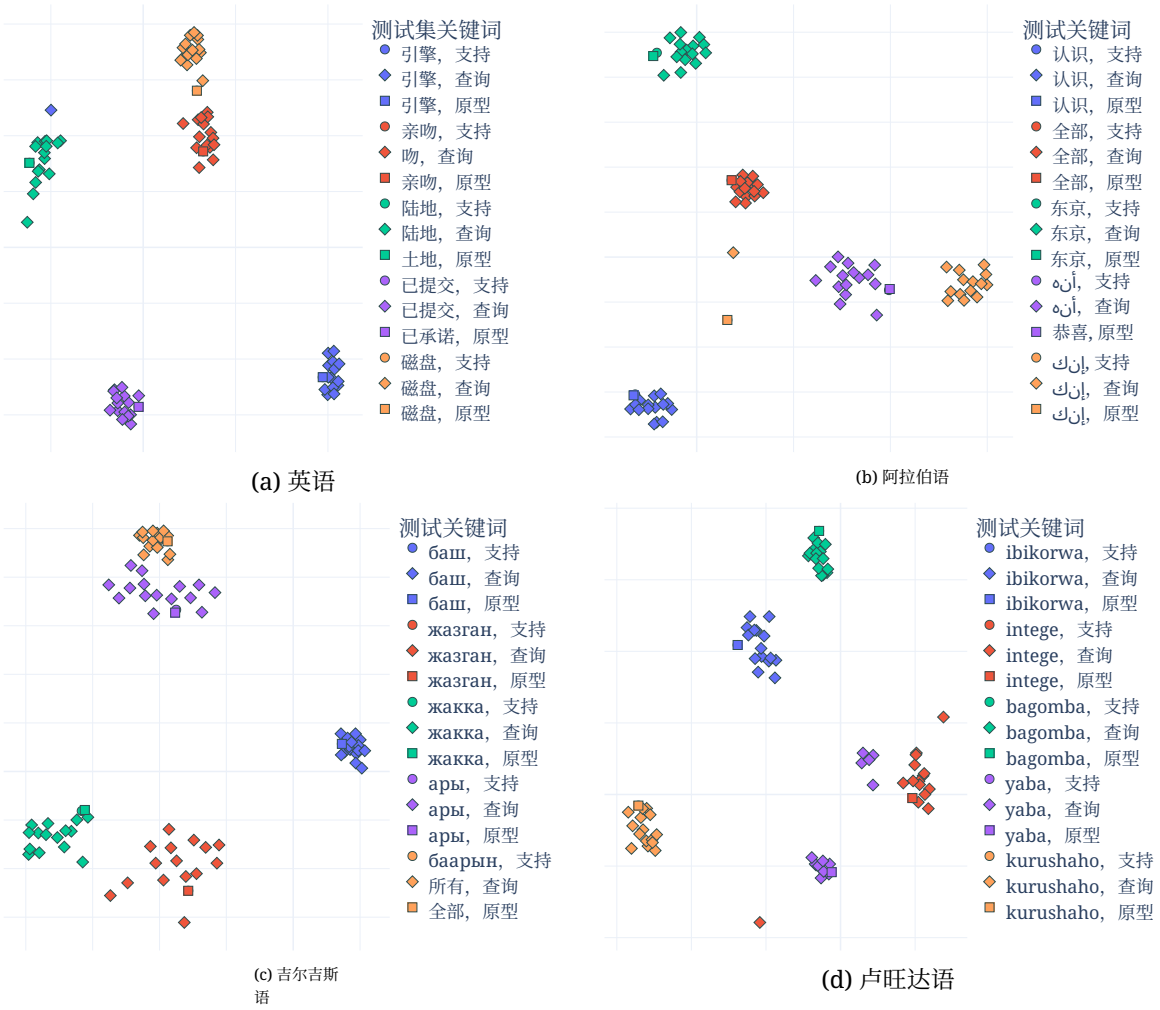



图 11. FLEURS 上的迁移评估 [7] 语言识别任务。

跨语言泛化：我们还研究了“base”多语言模型识别此前未见过的语言中的口语词的泛化能力，也就是说，这些语言并未用于模型训练。为此，我们选取了五种低资源语言：弗里西语（“fy-NL”）、蒙古语（“mn”）、马耳他语（“mt”）、斯洛文尼亚语（“sl”）和泰米尔语（“ta”）。我们剔除音频片段少于 15 条的关键词，并在具有挑战性的单样本设置下评估模型。无需针对这些新语言重新训练，模型开箱即能稳健运行。具体而言，除泰米尔语外，模型的准确率均超过 91%；在泰米尔语的 5 路 $\times 1$ 样本任务中，模型准确率达到 80%。这些结果表明，模型能够捕捉超越特定语言的语言特征。不过，泰米尔语准确率略有下降，这表明模型可能难以处理与训练语言差异较大的语言。仍需进一步研究，以提升模型对这类语言多样性较高的语言的泛化能力。

语言识别任务上的迁移评估：为确保完整性，我们还在 FLEURS [7] 语言识别任务的迁移设置下评估了以少样本方式训练的“基础”多语言模型，该任务总共涵盖 102 种语言。此任务要求根据一段音频片段识别所用语言。我们使用数据集中测试集的实例，并将每段音频裁剪或填充至 10 秒。我们不改变模型架构，也不执行任何训练步骤，以展示模型的泛化能力。我们在 3、5和10 分类设置下进行评估，支持样例数量分别为 5、10 和 20。我们发现，在 3 分类 $\times 10$样本任务中，模型的准确率超过 82%；在其余设置中，除 10 分类任务外，准确率均超过 73%。在这种情况下，准确率下降是意料之中的，因为在样例有限的情况下区分 10 种语言更加困难。总的来说，这些结果凸显了我们的多语言少样本方法在未见过的语言和口语词上，利用有限数据取得稳定表现的潜力。

使用 *t-SNE* 进行视觉表征分析：此外，在图 12 中，我们提供了一个在四种语言（即英语、阿拉伯语、韩语和卢旺达语）上训练的“基础”模型的 t-SNE [25] 嵌入，以展示基于不同口语词学习到的表征的解耦情况。我们使用 5 路 $\times 1$样本任务回合生成嵌入，并应用 t-SNE 生成 2D 投影。尽管每个类别仅提供一个支持样本作为示例，我们的模型仍然能够生成清晰分离的簇，分别对应不同语言中的各种口语词。这些结果进一步表明，我们的模型能够在测试时仅有有限标注数据的情况下，高效学习能够区分不同语言中口语词声学特征的表征，从而展现出在跨语言少样本识别方面的应用潜力。




(a) 英语

(b) 阿拉伯语

(c) 吉尔吉斯语

(d) 卢旺达语

图 12. 查询样例、支持集样例及按类别计算出的原型的 t-SNE 可视化结果。我们使用多语言“base”模型为 5-way × 1-shot 任务提取嵌入。为简洁起见，我们仅展示四种语言中单个任务回合的可视化结果。彩色显示并放大后效果更佳。

资源消耗分析： 对于第 2 节讨论的每个编码器（“small”和“base”），我们都分析其运行时资源需求，例如在三个嵌入式平台上的计算和能耗：NVIDIA Jetson Nano (4GB)、Raspberry Pi 4 和 Pixel 7（配置见图 13 ）。为简化配置，同时考虑到编码器模型是包含大量可学习参数的核心组件，我们仅使用编码器进行此项分析。我们根据这些平台在市场上的普及度，以及它们在边缘设备上运行机器学习工作负载的适用性来选择这些平台。

| | 设备 | 小型 | Base |
|---|---|---|---|
| 标准 | Raspberry Pi | 240.17 | 4480 |
| | Jetson Nano | 35.12 | 1954 |
| 量化 | 像素 | 5.12 | 78.07 |
| | 树莓派 | 6.92 | 107.13 |
| | Jetson Nano | 7.32 | 101.37 |

延迟（毫秒）

| | 设备 | 小型 | Base |
|---|---|---|---|
| 标准 | 树莓派 | 137.82 | 2011.44 |
| | Jetson Nano | 133.53 | 1227.24 |
| 量化 | 像素 | 3.55 | 39.71 |
| | Raspberry Pi | 4.15 | 71.75 |
| | Jetson Nano | 4.92 | 64.37 |

能耗（毫焦耳）

表 1. 所考察编码器 TinyNet-E 和 EfficientNet-v2 (M) 在不同设备上的平均能耗和延迟基准测试。为简化起见，计算这些指标时仅使用编码器网络，不包含其他组件（例如对数梅尔频谱图层）。Pixel 7 仅支持量化模型。

为了测试我们的系统，我们使用 PyTorch 基准测试模块对边缘设备[2] 和移动设备[3] 进行了评估，同时测试了两个版本的网络：全精度（float32）版本和针对移动设备优化的量化版本。由于 PyTorch 移动设备基准测试模块仅允许在设备上评估量化模型，因此我们仅在 Pixel 7 上使用量化版本进行实验。为了测量各平台的能耗，我们使用了外部设备（一台配备电流/功率监测 HAT 的 Raspberry Pi 4），通过采样电阻逐台监测设备的能耗，采样电阻为 0.1欧姆。对于 Pixel 7，我们停用了电池，以避免其干扰测量结果。表 1 展示了在所有三个平台上对深度模型进行 200 次推理运行所测得的平均运行时间（毫秒）和能耗（毫焦耳）。

我们注意到，我们的模型在 Raspberry Pi、Jetson Nano 和 Google Pixel 等各种硬件设备上都很高效。在 Raspberry Pi 上，标准 EfficientNet-v2 (M) 模型的延迟为 4480 毫秒，能耗为 2011.44 毫焦耳。“small”模型效率更高，延迟为 240.17 毫秒，能耗为 137.82 毫焦耳，但在泛化能力方面有所折衷。此外，对这些模型应用量化技术可显著提升效率。在 Raspberry Pi 上，量化后的“base”编码器延迟仅为 107.13 毫秒，能耗为 71.75 毫焦耳；与标准模型相比，分别提升了 19 倍和 44 倍。同样，Jetson Nano 上量化后的“small”编码器延迟为 7.32 毫秒，能耗为 4.92 毫焦耳；与标准模型相比，分别提升了 4 倍和 33 倍。总之，虽然我们的标准编码器能够在资源受限设备上高效运行，但量化能进一步提高速度和能效。这些优化后的模型使边缘硬件上的低功耗 AI 应用成为可能。我们的结果表明，即使在资源极其有限的计算平台上，也有望以计算上可行的方式部署少样本模型。

### 3.4 潜在应用与使用场景

我们相信，本文的贡献能够在许多领域产生重要影响。为此，我们在结尾列举了若干示例用法，以展示本系统在不同情境、平台和应用领域中的潜力。

**虚拟个人助理：**只需几个示例，我们的系统就能识别用户特定的新命令，从而让 Alexa 或 Siri 等虚拟个人助理变得更加实用、响应更加迅速。用户可以轻松教会助理理解新词或新短语，并根据自己的具体需求进行定制。

[2]https://pytorch.org/tutorials/recipes/recipes/benchmark.html [3] https://pytorch.org/tutorials/recipes/mobile_性能.html

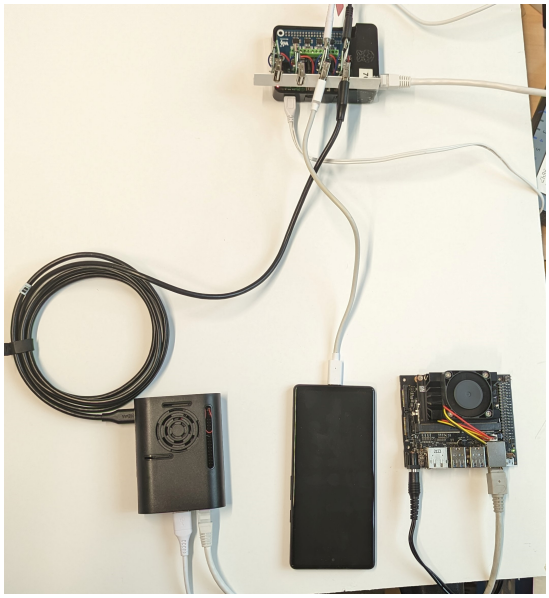


图13。嵌入式平台 用于估算我们模型在各种设备上的能耗和延迟消耗的平台设置 .

**智能家居自动化：** 少样本口语词识别器可用于通过语音命令控制智能物联网设备。只需少量示例即可识别新命令，用户便能轻松定制智能家居系统，使其识别并响应新的语音命令，让家居自动化更加顺畅。

**语言学习：** 少样本学习器可用于创建个性化语言学习工具，以适应每位学习者的个人需求。这些系统仅需极少量示例即可识别新词和短语，并为学习者提供即时反馈和发音建议，使语言学习更加高效有效。

**言语治疗：** 它还具有作为言语治疗有效工具的潜力，旨在提升言语障碍患者的语音能力。通过检测错误发音的能力，我们的系统可以提供及时反馈和支持，帮助患者改进发音并提升整体言语流利度。

**无障碍性：** 它还可用于改善残疾人士辅助应用的无障碍性，因为个性化定制仅需该个人提供少量示例，从而更容易根据具体的无障碍需求对系统进行个性化设置。

**信息学：** 在信息学领域，我们的系统可以提高在较长音频片段中记录和搜索特定词语或短语的效率。这项技术具有广泛的实际应用，可用于新闻业、执法或客户服务等众多领域；在这些领域中，快速查找和审阅音频录音中的特定信息对于自动化操作至关重要。

**游戏：** 为了创造更具沉浸感和互动性的游戏体验。玩家可以使用语音命令来控制游戏角色或做出决策。该系统仅凭少量支持样例就能识别新颖语音命令的能力，将使游戏过程更加个性化，从而提升用户的参与度。

## 4 相关工作

小样本学习领域已有大量研究，其目标是仅从少量示例中学习新概念。由于本文探索了用于语音词语检出的小样本学习方法，因此有必要回顾该领域的相关先前工作。小样本学习的主要方法提出了各种技术，旨在从有限数据中提取特征或学习更新规则，从而能够泛化到新概念。多年来，为此提出了若干方法，主要涉及元学习、生成建模和度量学习。基于元学习的方法旨在学习一种算法或模型，使其能够通过少量示例快速适应新任务。这包括学习如何学习分类器 [8], 学习如何将标签传播到未标记示例 [14], 以及通过记忆增强网络学习比较示例之间的相似度 [17]。生成式方法旨在为新类别生成合成但合理的示例，以扩充训练数据。这些方法包括学习生成图像 [18] 以及数据增强 [5, 19]。基于度量学习的方法 [21, 22] 学习距离度量，以对相似和不相似的示例进行分类，即便示例数量很少。在这一领域，开创性工作探索了基于原型的最近邻搜索 [21] 以及注意力机制 [26] ，使模型能够在学习新概念时利用先验知识。这些方法在基准小样本分类任务上取得了令人鼓舞的结果。鉴于基于距离的方法不仅在视觉领域，而且尤其在声音识别任务中都取得了巨大成功且简单易用 [6, 20, 27],我们采用原型网络 [21] 来开发小样本关键词检出系统。

早期关于口语词检测的研究探索了多种深度学习方法，包括卷积神经网络和循环神经网络等不同的神经网络架构 [3, 4, 15]。尽管这些方法实现了较小的模型规模，但其训练需要数千个目标关键词样本。其他针对低资源语言的关键词检出方法则采用多语言瓶颈特征 [13], 等技术，并利用语音合成 [10]。类似地， [16] 设计了一个涵盖 10 种语言的口语术语检索系统，以突破单语关键词检出的局限。近期， [11] 探索了迁移学习，通过使用极少量的标注数据对模型进行微调，以提高标签利用效率。相比之下，我们提出了一种小样本学习系统（PLiX），能够识别多种语言中的口语关键词，而无需通过重新训练来识别未见过的词语（类别）。我们构建了一个多语言声学模型，在元训练阶段学习跨语言的共享表征，而非采用迁移学习 [11], ，从而使模型能够泛化到可能出现的新语言和新关键词。此外，我们的方法具有通用性，因为它不需要特定语言的信息或资源（学习过程中仅使用原始音频），并且只需每个关键词提供少量支持样本即可实现出色的性能。据我们所知，所提出的 PLiX 系统是首个即插即用的小样本口语词检测系统，其训练数据包含数百万个音频片段，并且能够扩展到数千个关键词。与以往的方法相比，PLiX 在所支持的关键词数量和语言数量方面具有更强的可扩展性，使其能够应用于广泛的现实场景。

## 5 结论

我们介绍了 PLiX，这是一种端到端的少样本关键词检出系统，利用数百万个真实世界的音频片段学习深度模型，使其能够在测试时识别未见过的口语词（或类别）。实验结果表明，我们的方法在 20 种语言中检测数千个词时表现出色，展现了其仅凭单个支持样例便能泛化到新词的能力。开发能够将人类语言无缝融入新兴技术的系统，对于推进人机交互至关重要。因此，我们的工作是迈向构建实用、多语言关键词识别器的重要一步，这类识别器具有高度灵活性，并可根据用户需求进行个性化设置。总的来说，我们相信少样本元学习技术能够在加速机器理解人类语言这一新兴领域的发展中发挥关键作用。未来，我们计划通过整合上下文信息（例如说话者的年龄、性别、口音及其他

环境信息），以进一步提高准确率、探究隐私影响，并在更多语言和方言上评估该方法。

## 参考文献

[1] Rosana Ardila、Megan Branson、Kelly Davis、Michael Henretty、Michael Kohler、Josh Meyer、Reuben Morais、Lindsay Saunders、Francis M Tyers 和 Gregor Weber。2019。《Common Voice：大规模多语言语音语料库》。arXiv 预印本 arXiv:1912.06670（2019）。[2] Rich Caruana。1998。《多任务学习》。Springer。[3] Guoguo Chen、Carolina Parada 和 Georg Heigold。2014。使用深度神经网络实现小型关键词检出。载于《2014 IEEE 国际声学、语音与信号处理会议（ICASSP）》。IEEE，4087–4091。[4] Guoguo Chen、Carolina Parada 和 Tara N Sainath。2015。使用长短期记忆网络实现基于查询示例的关键词检出。载于《2015 IEEE 国际声学、语音与信号处理会议（ICASSP）》。IEEE，5236–5240。[5] Zitian Chen、Yanwei Fu、Yu-Xiong Wang、Lin Ma、Wei Liu 和 Martial Hebert。2019。用于单样本学习的图像变形元网络。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》。8680–8689。[6] Szu-Yu Chou、Kai-Hsiang Cheng、Jyh-Shing Roger Jang 和 Yi-Hsuan Yang。2019。利用注意力相似度学习匹配瞬态声音事件，以实现小样本声音识别。载于《ICASSP 2019—2019 IEEE 国际声学、语音与信号处理会议（ICASSP）》。IEEE，26–30。[7] Alexis Conneau、Min Ma、Simran Khanuja、Yu Zhang、Vera Axelrod、Siddharth Dalmia、Jason Riesa、Clara Rivera 和 Ankur Bapna。2022。FLEURS：语音通用表征的小样本学习评估。arXiv 预印本 arXiv:2205.12446（2022）。https: //arxiv.org/abs/2205 12446 。Chelsea Finn、Pieter Abbeel 和 Sergey Levine。2017。用于深度网络快速适应的模型无关元学习。载于《国际机器学习会议》。PMLR，1126–1135。[9] Kai Han、Yunhe Wang、Qiulin Zhang、Wei Zhang、Chunjing Xu 和 Tong Zhang。2020。模型魔方：调整 TinyNet 的分辨率、深度和宽度。《神经信息处理系统进展》33（2020），19353–19364。[10] James Lin、Kevin Kilgour、Dominik Roblek 和 Matthew Sharifi。2020。利用有限的合成语音数据训练关键词检出器。载于《ICASSP 2020—2020 IEEE 国际声学、语音与信号处理会议（ICASSP）》。IEEE，7474–7478。[11] Mark Mazumder、Colby Banbury、Josh Meyer、Pete Warden 和 Vijay Janapa Reddi。2021。面向任意语言的小样本关键词检出。arXiv 预印本 arXiv:2104.01454（2021）。[12] Mark Mazumder、Sharad Chitlangia、Colby Banbury、Yiping Kang、Juan Manuel Ciro、Keith Achorn、Daniel Galvez、Mark Sabini、Peter Mattson、David Kanter 等。2021。多语言口语词语料库。载于《第三十五届神经信息处理系统会议数据集与基准测试专题（第二轮）》。[13] Raghav Menon、Herman Kamper、Ewald Van Der Westhuizen、John Quinn 和 Thomas Niesler。2018。利用多语言瓶颈特征提取器和对应自编码器，为几乎零资源、无需自动语音识别的关键词检出探索特征。arXiv 预印本 arXiv:1811.08284（2018）。[14] Sachin Ravi 和 Hugo Larochelle。2017。将优化视为小样本学习的模型。载于《国际学习表征会议》。[15] Tara Sainath 和 Carolina Parada。2015。用于小型关键词检出的卷积神经网络。（2015）。[16] Nay San、Martijn Bartelds、Mitchell Browne、Lily Clifford、Fiona Gibson、John Mansfield、David Nash、Jane Simpson、Myfany Turpin、Maria Vollmer 等。2021。利用神经表征，促进获取濒危语言的未转写语音。arXiv 预印本 arXiv:2103.14583（2021）。[17] Adam Santoro、Sergey Bartunov、Matthew Botvinick、Daan Wierstra 和 Timothy Lillicrap。2016。利用记忆增强型神经网络进行元学习。载于《国际机器学习会议》。PMLR，1842–1850。[18] Eli Schwartz、Leonid Karlinsky、Joseph Shtok、Sivan Harary、Mattias Marder、Abhishek Kumar、Rogerio Feris、Raja Giryes 和 Alex Bronstein。2018。一种用于小样本目标识别的有效样本合成方法。载于《神经信息处理系统年会》。神经信息处理系统基金会。[19] Eli Schwartz、Leonid Karlinsky、Joseph Shtok、Sivan Harary、Mattias Marder、Abhishek Kumar、Rogerio Feris、Raja Giryes 和 Alex Bronstein。2018。Delta 编码器：一种用于小样本目标识别的有效样本合成方法。《神经信息处理系统进展》31（2018）。[20] Bowen Shi、Ming Sun、Krishna C Puvvada、Chieh-Chi Kao、Spyros Matsoukas 和 Chao Wang。2020。通过元学习实现小样本声学事件检测。载于《ICASSP 2020—2020 IEEE 国际声学、语音与信号处理会议（ICASSP）》。IEEE，76–80。[21] Jake Snell、Kevin Swersky 和 Richard Zemel。2017。用于小样本学习的原型网络。《神经信息处理系统进展》30（2017）。[22] Flood Sung、Yongxin Yang、Li Zhang、Tao Xiang、Philip HS Torr 和 Timothy M Hospedales。2018。学习比较：用于小样本学习的关系网络。载于《IEEE 计算机视觉与模式识别会议论文集》。1199–1208。

[23] Mingxing Tan 和 Quoc Le。2021。EfficientNetV2：更小的模型与更快的训练。载于国际机器学习会议。PMLR，10096–10106。[24] Sebastian Thrun。1998。《终身学习算法》。学习如何学习 8（1998），181–209。[25] Laurens Van der Maaten 和 Geoffrey Hinton。2008。使用 t-SNE 可视化数据。机器学习研究期刊 9，11（2008）。[26] Oriol Vinyals、Charles Blundell、Timothy Lillicrap、Daan Wierstra 等。2016。用于单样本学习的匹配网络。神经信息处理系统进展 29（2016）。[27] Yu Wang、Justin Salamon、Nicholas J Bryan 和 Juan Pablo Bello。2020。小样本声音事件检测。载于 *ICASSP 2020—2020 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，81–85。