arXiv:2304.03416v1 [eess.SP] 6 Apr 2023

# 唤醒还是不唤醒：通过逐步细化减少关键词误报

*Yashas Malur Saidutta*, Rakshith Sharma Srinivasa*, Ching-HuaLee, Chouchang Yang, Yilin Shen,HongxiaJin*三星美国研究院，美国加利福尼亚州山景城

**摘要**

关键词检测系统会持续处理音频流，以检测关键词。设计这类系统时，最具挑战性的任务之一是减少误报（误报），即系统在关键词实际上并未被说出时，却错误地识别出关键词。本文提出了一种简单而巧妙的解决方案，其依据是全概率定律。我们证明，现有的深度关键词检测机制可以通过逐步细化得到改进：系统首先判断输入音频是否为语音，然后判断输入是否类似关键词，最后识别说出的是哪个关键词。我们在参数量从 13K 到 2.41M 的多种模型上进行了验证，结果表明，逐步细化技术可将域内留出误报数据上的误报率最多降低至原来的 1/8，并将域外（OOD）误报数据上的误报率最多降低至原来的 1/7。此外，我们提出的方法即插即用，可应用于任何深度关键词检测模型。

索引词— 深度关键词检测、误报

## 1. 引言

从移动设备到家用电器，关键词检测（KWS）是人机交互的基石之一 [1]。KWS 系统会持续处理音频流，以检测关键词。然而，大部分时间里，KWS 系统处理的只是包含非关键词语音或非语音（也称为噪声）的音频，因为关键词只会偶尔出现。设想一个始终开启的 KWS 系统，其误报（FA）率为 5%，每 0.1 秒处理一秒音频（由于与之前的音频帧存在重叠）。这样每小时就会产生 *1800* 次误报。更重要的是，KWS 系统一旦检测到关键词，就会触发其他规模更大的云端系统，例如自动语音识别（ASR）。因此，KWS 系统中的误报可能会引发隐私问题，因为用户音频会在不必要的情况下被录制并上传到云端。最后，触发其他下游任务还会导致不必要的功耗，这对于智能手机和智能手表等电池供电设备而言并不可取。

以往采用深度神经网络（DNN）的 KWS 工作，即深度 KWS 系统，取得了令人瞩目的成果 [2]。然而，这些系统将非关键词语音和非语音音频视为两个互斥的类别，且认为它们与关键词集合的距离相等。图 1（a）展示了单关键词系统中的这一观点。但实际上，语音是所有音频的一个子集，而关键词是语音的一个子集，如图 1（b）所示。所提出的观点使我们能够利用这样一个事实：关键词与非关键词语音有更多相似之处，



(a) 现有技术视角　　(b) 提议的观点

**图1**：此处，红色表示非语音，绿色表示非关键词语音，蓝色表示关键词。（a）以往工作将这些集合视为彼此互斥的类别等距。（b）在所提出的观点中，所有音频都由点线界定。语音成分构成所有音频中的一个更小子集。关键词构成语音集合中的一个更小子集。

因为两者都是语音。对于关键词，也可以作出类似的论证。关键词作为独立命令说出，因此其能量会随着话语呈现出特征性的起伏。相反，普通语音的能量曲线则更加平稳。

基于上述见解，我们提出了一种基于逐步细化的深度关键词检索分类器。该分类器首先判断输入音频是否为语音。如果是语音，则判断该语音是否类似关键词或属于非关键词语音。最后，如果判定为类似关键词，最终分类器会识别出所说的是哪个关键词。重要的是，这种方法并不试图部署三个独立的分类器模型。相反，我们复用骨干模型的内部表示来构建三个不同的分类分支，其中每个分支都通过复制模型最后一层或至多两层来实现。这种分层监督能够带来更丰富的学习效果，并支持训练误报更低的模型。我们在下文总结我们的贡献：

- 我们提出了一种通用方法，用于降低深度关键词检索模型中的误报率。该方法可用于任意选定的架构，同时仅需 negligible 的内存需求和计算量增加。
- 在最先进（SOTA）的深度关键词检索模型中，引入逐步细化可使保留的误报数据上的误报率最多降低至原来的八分之一。
- 逐步细化 在分布外误报数据上表现出持续改进，最高可达 7倍。在分布外数据上的性能至关重要，因为训练期间不可能让模型接触所有可能的音频数据。

## 2. 相关工作

已有许多关键词检索研究采用深度学习技术，取得了优于传统基于模型的方法的性能，例如 [3–16]。深度关键词检索研究尝试从原始音频中学习参数化特征提取 [17] ，或学习架构本身 [18, 19]。深度关键词检索系统的另一个重要方面是模型大小。由于关键词检索模型通常部署在设备上，并持续处理音频流，因此其内存和计算开销需要较低。已有研究。

* 贡献相同。

©2023 IEEE。允许个人使用本材料。任何其他用途均须获得 IEEE 的许可，包括在任何当前或未来媒体中重印/再版本材料以用于广告或宣传目的、创作新的汇编作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护部分。DOI:

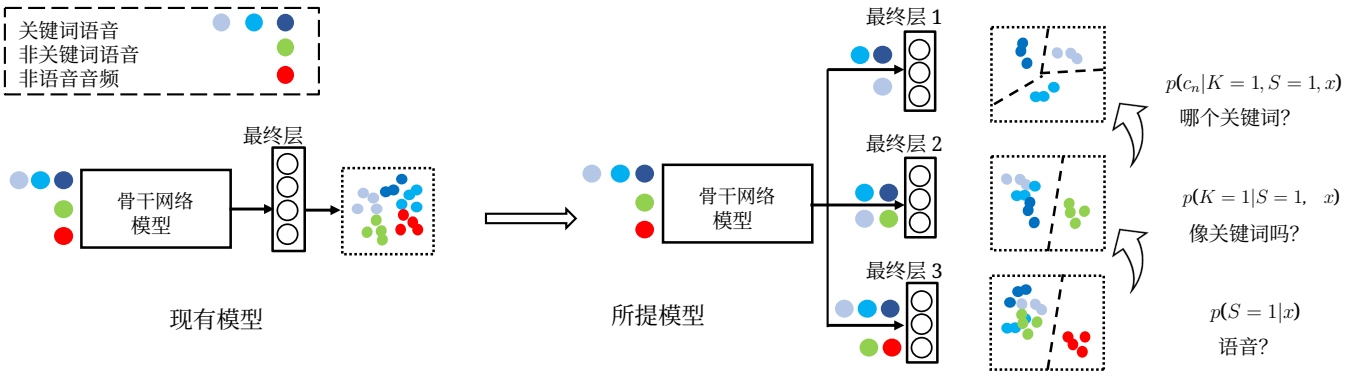



**图 2**：系统概述：本文提出了一种减少关键词检索中误报的新方法。我们将关键词检测问题视为多阶段分类问题，并采用逐步细化来滤除非关键词音频。

已有几篇论文专注于构建满足这一标准的关键词检索模型 [10,20,21]。关于深度关键词检索系统的全面综述，读者可参考 [2]。

与上述研究不同，深度关键词检索的另一个关键方面是减少误报。出于隐私和功耗方面的考虑，误报率应尽可能低，正如第 1 节所述。 [22], 中设计了一种高效的误报缓解系统，但它们需要在话语结束后获取音频，以获得上下文信息。最近， [23], 中的作者指出，良好的非语音隐马尔可夫模型对于唤醒词系统的良好性能至关重要。本文的动机与此类似，即利用性能良好的语音与非语音深度分类器，以及关键词与非关键词深度分类器，以提升性能。所提出的方法与上述研究互为补充，可用于增强其中任何一种方法。

### 3. 使用逐步细化降低误报

我们方法的核心思想基于以下观察：现有的关键词检测模型仅将关键词语音、非关键词语音和非语音音频视为属于不同类别。基于这一假设，现有模型将关键词检测视为一个多类分类问题。然而，如第 1 节所述，这些不同类型的音频具有更自然的层次结构。

上述层次结构可以使用如下所示的全概率定律进行量化。考虑一个词汇表中含有 $N$ 个词的关键词检测系统。令 $c_n$ 表示第 $n$ 个关键词。首先，我们定义随机变量 $S$：当音频帧 $x$ 中包含语音时取值为 1，否则取值为 0。我们还定义另一个随机变量 $K$：当 $x$ 中包含关键词语音时取值为 1，当其包含非关键词语音时取值为 0。于是，对于给定的输入音频 $x$，有

$$p(c_n|x) = \text{Probability of } x \text{ being keyword } c_n \tag{1}$$

$$= \sum_{k,s\in\{0,1\}} p_{C|K,S,X}(c_n|k,s,x)p_{K|S,X}(k|s,x)p_{S|X}(s|x) \tag{2}$$

$$= p(c_n|k=1,s=1,x)p(k=1|s=1,x)p(s=1|x), \tag{3}$$

其中，(2) 是由音频类别的层次结构性质（图 1(b)）得出的；(3) 则是因为 $p(c_n) = 0$ if $K = 0$or $S = 0$。

我们利用上述输出概率的层级结构来设计模型。图 2 展示了所提出的关键词检测系统的示意图。如图所示，我们的架构由一个骨干神经网络和三个分支组成：关键词分类、关键词分支以及语音分支。所有分支最多由两层组成，其中关键词分支和语音分支使用sigmoid函数激活，关键词分类分支使用softmax激活。语音分支经过训练，用于将给定输入分类为人类语音或其他情况；关键词分支经过训练，用于将给定输入分类为关键词或其他情况。最后，关键词分类分支经过训练，用于分类给定输入中出现的是哪个关键词。如下所述，这些分支在训练和推理期间的使用方式有所不同。

**训练：** 在训练过程中，我们使用（3）对数据流经网络的过程进行建模。骨干模型首先为整个小批次生成中间嵌入表示。然后，仅将对应于包含关键词的输入的嵌入表示传入关键词分支（图2中最上方的分支），以将其分类为某个关键词。这直接对概率 p $(c_n|K=1,\ S=1,\ x)$ 进行建模。类似地，为了对概率 p $(K=1|S=1,\ x)$ 进行建模，我们将对应于包含语音的输入的嵌入表示传入类似关键词的分支（图2中的中间分支）。最后，所有嵌入表示都经过语音分支（图2中的底部分支），以对概率 p $(S=1|x)$ 进行建模。随后使用这三个分支的输出计算三个损失：用于关键词分类的 softmax 损失，以及用于关键词分支和语音分支二分类的加权焦点损失 [24]。最终的损失函数为

$$\mathcal{L} = \mathcal{L}_{\text{softmax}} + \lambda_1\mathcal{L}_{\text{keyword branch}} + \lambda_2\mathcal{L}_{\text{speech branch}}, \tag{4}$$

其中 $\lambda_1$, $\lambda_2$ 表示损失函数各组成部分的相对权重。这些损失函数中的不同类别按照该类别样本数量的倒数进行加权。在反向传播过程中，每个分支仅从其对应的损失组成部分接收梯度（例如，最顶部的分支仅从 $\mathcal{L}_{\text{softmax}}$ 接收梯度）。只有骨干模型从全部三个损失组成部分接收梯度。

**数据池化带来的性能提升：** 所提出模型中的层级结构自然地使语音分支和关键词分支能够进行数据池化。对于语音分支，所有关键词语音和非关键词语音音频都被池化为一个类别，从而扩大训练数据规模。这也使我们能够使用更大的非语音数据集，而不会造成类别不平衡。同样，各种关键词对应的数据也会在关键词分支中池化为一个类别。

**推理：** 在对音频 $x$进行推理时，关键词分类分支会给出一个在关键词集合 $\{1,\ldots,N\}$ 上的多项分布，其形式为 $p(c_n|K=1,S=1,x) := p_{c_n}$,

关键词分支给出音频是否类似关键词的分布 $p(K = i|S = 1, x)$：$= p_{K=i}$，而语音分支给出音频是否为语音的分布 $p(S = i|x)$：$= p_{S=i}$，其中 $i \in \{0, 1\}$。我们将这些分布组合起来，得到形式为 $\mathbf{p}$：$[p_{c_1} p_{K=1} p_{S=1}, \ldots, p_{c_N} p_{K=1} p_{S=1}, p_{K=0} p_{S=1}, p_{S=0}]$ 的多项分布。利用 $\mathbf{p}$，我们计算 $\hat{l} = \arg\max_{l \in \{1,\ldots,N+2\}} \mathbf{p}[l]$。如果 $\hat{l} = n$，其中 $n \in \{1, \ldots, N\}$，则检测到 $n^{th}$ 关键词。如果 $\hat{l} = N + 1$，则检测到非关键词语音；如果 $l = N + 2$，则检测到非语音。

## 4. 实验结果

**数据集：** 我们评估了所提方法在 Google Speech Commands 数据集 V2（GSC-V2） [26]上的性能。我们针对两个变体进行实验：一种仅使用 [26] 中的十个关键词，另一种使用全部三十五个关键词。对于这些类别，我们使用 [26]提供的相同训练集、验证集和测试集划分。对于“非关键词语音”（在 [26]中称为“未知”）类别，我们使用来自 childspeech 数据集 [27], 的 760 个一秒样本，以及来自 AVSpeech 数据集 [28]的 4565 个一秒样本。对于“非语音”（在 [26] 中称为“静音”）类别，我们使用来自 Audioset [29]的 7493 个一秒样本。由于 Audioset 中的一些样本包含语音，我们使用语音活动检测器 [30]将其过滤掉。所有这些样本均按 8:1:1 的比例划分为训练集、验证集和测试集。

我们还对训练好的模型进行域外（OOD）实验，以测试模型在未见条件下对非关键词语音和非语音的鲁棒性。为此，我们使用 MUSAN 数据集 [31]。每个语音文件采样得到三个 1s 帧，每个噪声文件采样得到一个 1s 帧，共计 2155 个样本用于误报测试。

**实现细节：** 我们在多种模型上进行实验，包括 BC-ResNet 模型 [10], 关键词变换器 [13], 和共享权重自注意力时延神经网络（TDNN-SWSA，简称 TSWSA） [14]。选择这些模型的动机是同时考察最先进的模型（ [10, 13]）和小型模型（ [14]）。

***BC-ResNet***：我们修改了原始架构，在最终 softmax层 [10]之前添加一个包含 80 个神经元的线性层，并接入 ReLU。为进行逐步细化，我们将 80 神经元层替换为三个 32 神经元层（每个分支各一个），并为它们分别设置输出层。我们使用随机梯度下降和单周期学习率调度器 [32]训练模型 25 个训练轮次。学习率在七个训练轮次内从 0.004 增至 0.1，随后衰减至 $4 \times 10^{-6}$。小批量大小为 100。根据 [32], 动量在 0.85 至 0.95 之间变化。数据增强等其他设置均遵循 [10]。

关键词变换器（***KWT***）：对于基线KWT模型，我们采用 [13] 中指定的相同架构。我们还使用 [13] 中描述的相同数据增强方案。为纳入所提出的方法，我们使用倒数第二层的输出。关键词分类器由层归一化和一个线性层组成（与 [13] 中相同）。语音分类器和类关键词分类器各由一个大小为 $n \times n$ 的线性层组成，其中 $n$ 为嵌入维度，后接 ReLU，再接一个大小为 $n \times 1$ 的线性层。我们对关键词分类器使用 $N$ 类交叉熵损失，对另外两个分支使用二元交叉熵损失（经过 sigmoid 函数非线性变换后）。我们仅提供 KWT-1 和 KWT-2 模型的结果，因为 KWT-3 的基线模型在训练期间未能很好地收敛。我们的实现代码来源于 https://github.com/ID56/Torch-KWT 提供的 PyTorch 实现。

***TSWSA***：我们对原始架构进行了修改，在最终的 softmax 输出层 [14] 前添加了一个包含 32 个神经元的线性层，后接 ReLU。为实现逐步细化，我们添加了三个各含 10 个神经元的线性层（以及输出层）。模型优化器和调度器与 BC-ResNet 中所述相同。我们发现，100 个样本的迷你批次优于原来的 32 个。其他所有设置（如梅尔频率倒谱系数）均遵循 [14]。

**性能指标：** 我们计算准确率和加权F1分数，以衡量关键词检索的整体性能。该指标在$N + 2$ 个类别上计算，包括 $N$ 个关键词、非关键词语音和非语音类别。我们将误报计算为被分类为关键词的非关键词语音和非语音音频所占的比例。

### 4.1. 结果

我们比较了本方法（逐步细化，简称*SR*）在分类准确率、F1 分数以及域内未见误报数据和分布外误报数据上的误报率方面带来的提升。表1展示了关键词数量为十个时的结果。从结果可以看出，**加入逐步细化后，所有模型在域内和分布外误报数据上的误报表现均得到持续改善**，同时准确率与基线模型基本持平或更高。此外，**误报表现的提升并未大幅增加内存（参数规模）或计算量（乘累加运算次数）**。在表2中，我们使用 BC-ResNet 模型对包含三十五个关键词的数据进行了实验；这些模型在十个关键词的数据上表现最佳。我们发现，**随着关键词数量增加，逐步细化对误报表现的改善作用更大**。

**表1**：在 GSC-V2 数据集上进行训练的结果，包含十个关键词、一个非关键词语音类别（来自 AVSpeech 和 Child-speech）以及一个非语音类别（来自 Audioset）。FA（OOD）使用来自 MUSAN 数据集的语音和噪声计算得出。逐步细化 指所提出的逐步细化方法。

| 模型 | 准确率（%） | F1 | 误报（%） | 分布外误报 (%) | #参数量 | #乘累加运算次数 |
|---|---|---|---|---|---|---|
| BCR-1 [10] | 93.42 | 0.965 | 4.10 | 6.08 | 13.0k | 3.34M |
| BCR-1+SR | 93.22 | 0.964 | **1.41** | **2.13** | 13.0k | 3.34M |
| BCR-1.5 [10] | 94.05 | 0.967 | 3.99 | 6.45 | 22.3k | 5.89M |
| BCR-1.5+SR | 94.24 | 0.969 | **1.17** | **2.65** | 22.5k | 589万 |
| BCR-2 [10] | 94.17 | 0.969 | 4.23 | 8.03 | 33.8k | 9.03M |
| BCR-2+SR | 95.26 | 0.975 | **1.11** | **2.22** | 34.3k | 9.03M |
| BCR-3 [10] | 95.48 | 0.976 | 3.79 | 5.89 | 63.5k | 17.1M |
| BCR-3+SR | 95.77 | 0.978 | **1.01** | **2.13** | 64.5k | 17.1M |
| BCR-6 [10] | 96.48 | 0.982 | 2.25 | 4.36 | 205k | 5530万 |
| BCR-6+SR | 96.27 | 0.981 | **1.04** | **1.58** | 208k | 5530万 |
| BCR-8 [10] | 96.70 | 0.983 | 2.15 | 4.22 | 344k | 9260万 |
| BCR-8+SR | 96.82 | 0.984 | **0.77** | **1.72** | 348k | 92.6M |
| KWT-1 [13] | 94.9 | 0.968 | 2.5 | 7.19 | 551k | 56.40M |
| KWT-2+SR | 94.48 | 0.965 | **2.3** | **5.38** | 559.7k | 56.42M |
| KWT-2 [13] | 94.62 | 0.967 | 2.75 | 5.75 | 2,381k | 243.2M |
| KWT-2+SR | 94.76 | 0.967 | **2.08** | **5.38** | 2,415k | 243.2M |
| TSWA [14] | 91.56 | 0.955 | 4.91 | 9.84 | 12.8K | 0.36M |
| TSWSA+SR | 91.21 | 0.953 | **1.78** | **2.97** | 12.5K | 0.36M |

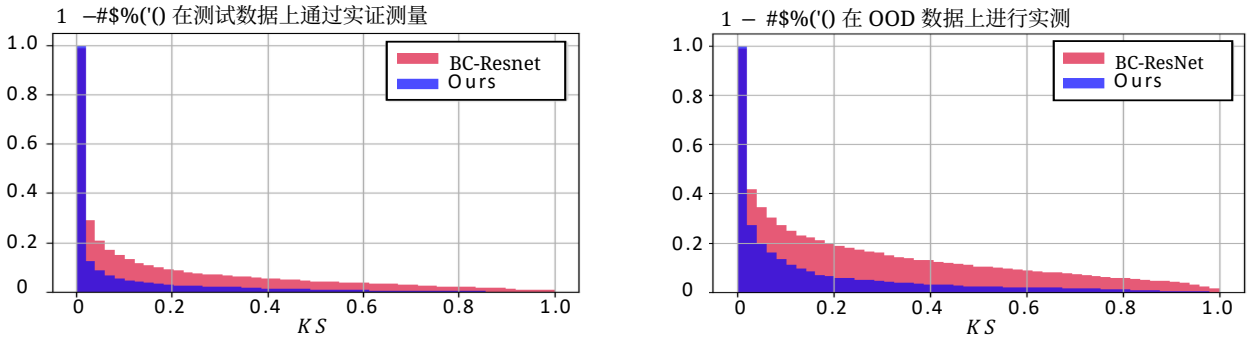
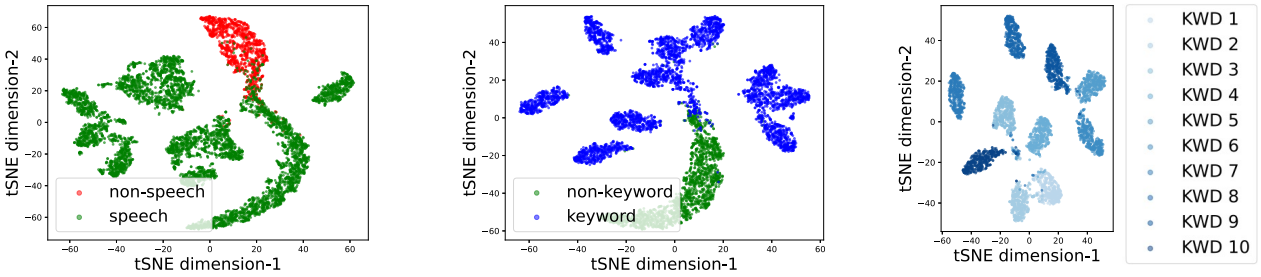



**图 3**：我们展示了在带有和不带有 逐步细化 的BCResNet-2 上，留出测试数据（左）和 OOD 数据（右）中非关键词、非语音部分的随机变量 $KS$ 的逆累积分布函数（cdf）。对于较低的误报，$KS$ 的分布应集中在 0附近。在上述图中，我们观察到，对于留出测试数据和 OOD 数据，$p_{\text{baseline}}(KS < \alpha) > p_{\text{ours}}(KS < \alpha)$ 在大多数 $\alpha \in [0, 1]$ 中均如此。由于逐步细化具有更强的判别能力，这种效果在 OOD 数据上更加明显。




(a) 语音与非语音分支　　(b) 关键词与非关键词分支　　（c）所有关键词的关键词分类器

**图4**：使用 tSNE [25] 可视化三个不同分支学习到的嵌入表示。

**表2**：在包含三十五个关键词的 GSC-V2 数据集上进行训练的结果。*SR* 指提出的逐步细化方法。

| 模型 | 准确率 (%) | F1 | 误报（%） | 误报 OOD (%) | #参数 | #乘累加运算次数 |
|---|---|---|---|---|---|---|
| BCR-1 [10] | 90.17 | 0.944 | 13.49 | 14.06 | 15.1k | 3.34M |
| BCR-1+SR | 89.65 | 0.942 | **1.61** | **1.95** | 13.8K | 3.34M |
| BCR-1.5 [10] | 91.61 | 0.953 | 12.02 | 15.41 | 24.4k | 5.89M |
| BCR-1.5+SR | 91.38 | 0.952 | **1.58** | **2.55** | 23.4K | 589 万 |
| BCR-2 [10] | 92.58 | 0.959 | 10.47 | 13.36 | 35.9k | 9.03M |
| BCR-2+SR | 93.00 | 0.962 | **1.98** | **3.62** | 35.1K | 903万 |
| BCR-3 [10] | 93.73 | 0.965 | 8.69 | 11.69 | 65.5k | 1710 万 |
| BCR-3+SR | 94.23 | 0.968 | **1.61** | **2.83** | 65.3K | 17.1M |
| BCR-6 [10] | 95.10 | 0.973 | 6.71 | 9.47 | 207k | 55.3M |
| BCR-6+SR | 95.54 | 0.976 | **1.21** | **1.58** | 209K | 55.3M |
| BCR-8 [10] | 95.39 | 0.975 | 6.04 | 8.86 | 346k | 92.6M |
| BCR-8+SR | 95.56 | 0.976 | **0.91** | **1.76** | 349K | 92.6M |

为进一步研究本方法对误报率的影响，我们在图3中展示了 $KS$（一个随机变量，用于表示给定音频输入是否同时为语音且属于某个关键词）的累积分布。请注意，$KS$ 是任意关键词概率的最大可能值。因此，$KS$ 的上界较小，表明误报率较低。对于基线模型，我们根据经验估计 $KS$，其形式为 $KS = 1 - \sum_{i=1}^{N} p(c_i)$，其中 $c_i$ 是第 $i$个关键词。我们的方法（以蓝色显示）始终得到更低的 $KS$，从而带来**更好的最坏情况误报率**。这一点也体现在我们的表 1 和表 2 所示的误报率。

我们提出的方法将问题分解为多个步骤：首先判断输入音频是否为语音；如果是语音，则判断其是否类似关键词；最后，如果类似关键词，则确定说出的是哪个关键词。这有助于系统学习专门的嵌入表示，从而更容易完成这些子任务。在图 4 中，我们使用 tSNE [25] 对这些嵌入表示进行了可视化。在图 4( a) 中，语音分支只专注于将语音与非语音区分开，而不是将不同的语音类型聚类为不同的簇。然而，图 4(b) 所示的后续关键词分支则将语音类型区分为类似关键词的语音和一般语音。最后，如图 4(c) 所示，关键词分类器将不同的关键词聚类到各自的簇中。因此，逐步细化使不同分支能够针对更简单的子问题学习专门的嵌入表示，从而获得更强的判别能力。

## 5. 结论

现有的关键词检索系统将非关键词语音和非语音视为与关键词类别距离相等的两类。然而，实际上，非关键词语音与关键词彼此之间比它们与非语音更加相似。我们利用这一洞见构建了一个逐步细化的分类器，用于判断输入音频是否为语音；如果是语音，则判断其属于类似关键词的语音还是普通语音；最后，如果属于类似关键词的语音，则判断说出的是哪个关键词。这也可视为对全概率公式一种简单而精妙的扩展。我们在参数量从 13K 到 2.41M 的各种规模模型上进行了实验，并表明该

该模型的检测准确率与基线相近，同时在域内误报数据上误报最多减少至原来的八分之一，在分布外误报数据上最多减少至原来的七分之一。

### 6. 参考文献

[1] R. Haeb-Umbach、S. Watanabe、T. Nakatani、M. Bacchiani、B. Hoffmeister、M. L. Seltzer、H. Zen 和 M. Souden，《数字家庭助理的语音处理：将信号处理与深度学习技术相结合》，*IEEE SignalProcessing Magazine*，第 36 卷，第 6 期，第 111–124 页，2019年。[2] I. L´opez-Espejo、Z.-H. Tan、J. Hansen 和 J. Jensen，《深度语音关键词检测：综述》，*IEEEAccess*，2021年。[3] M. Sun、A. Raju、G. Tucker、S. Panchapagesan、G. Fu、A. Mandal、S. Matsoukas、N. Strom 和 S. Vitaladevuni，《用于小型关键词检测的长短期记忆网络最大池化损失训练》，发表于*IEEE Spoken Language Technology Workshop (SLT)*，2016年，第 474–480 页。[4] S. Arık、M. Kliegl、R. Child、J. Hestness、A. Gibiansky、C. Fougner、R. Prenger 和 A. Coates，《用于小型关键词检测的卷积循环神经网络》，发表于国际语音通信协会年会 *(Interspeech)*，2017年，第 1606–1610 页。[5] T. Sainath 和 C. Parada，《用于小型关键词检测的卷积神经网络》，发表于国际语音通信协会年会 *(Interspeech)*，2015年，第 1478–1482 页。[6] R. Tang 和 J. Lin，《用于小型关键词检测的深度残差学习》，发表于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2018年，第 5484–5488 页。[7] S. Choi、S. Seo、B. Shin、H. Byun、M. Kersner、B. Kim、D. Kim 和 S. Ha，《移动设备实时关键词检测的时序卷积》，发表于国际语音通信协会年会 *(Interspeech)*，2019年，第 3372–3376 页。[8] X. Li、X. Wei 和 X. Qin，《采用多尺度时序卷积的小型关键词检测》，发表于国际语音通信协会年会 *(Interspeech)*，2020年，第 1987–1991 页。[9] M. Xu 和 X.-L. Zhang，《采用压缩激励模块的深度可分离卷积 ResNet，用于小型关键词检测》，发表于国际语音通信协会年会 *(Interspeech)*，2020年，第 2547–2551 页。[10] B. Kim、S. Chang、J. Lee 和 D. Sung，《面向高效关键词检测的广播式残差学习》，*arXiv*预印本 *arXiv:2106 04140*，2021年.。[11] M. Zeng 和 N. Xiao，《 DenseNet 与 BiLSTM 在关键词检测中的有效结合》，*IEEE Access*，第 7 卷，10775，2019年 10767pp. – 。[12] O. Rybakov、N. Kononenko、N. Subrahmanya、M. Visontai 和 S. Lau-renzo，《移动设备上的流式关键词检测》，发表于国际语音通信协会年会 *(Interspeech)*，2020年，第 2277–2281 页。[13] A. Berg、M. O' Connor 和 M. T. Cruz，《关键词变换器：一种用于关键词检测的自注意力模型》，发表于国际语音通信协会年会 *(Interspeech)*，2021年，第 4249–4253 页。[14] Y. Bai、J. Yi、J. Tao、Z. Wen、Z. Tian、C. Zhao 和 C. Fan，《一种采用共享权重自注意力的小型时延神经网络关键词检测模型》，发表于国际语音通信协会年会 *(Interspeech)*，2019年，第 2190–2194 页。

[15] D. C. De Andrade、S. Leo、M. L. Da Silva Viana 和 C. Bernkopf，《用于语音命令识别的神经注意力模型》，*arXiv*预印本 *arXiv:1808.08929*，2018年。[16] X. Chen、S. Yin、D. Song、P. Ouyang、L. Liu 和 S. Wei，《采用图卷积网络的小型关键词检测》，发表于*IEEE*自动语音识别与理解研讨会 *(ASRU)*，2019年，第 539–546 页。[17] S. Mittermaier、L. K¨urzinger、B. Waschneck 和 G. Rigoll，《利用原始音频数据和 sinc 卷积进行小型关键词检测》，发表于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2020年，第 7454–7458 页。[18] B. Zhang、W. Li、Q. Li、W. Zhuang、X. Chu 和 Y. Wang，《AutoKWS：采用可微架构搜索的关键词检测》，发表于*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2021年，第 2830–2834 页。[19] T. Mo、Y. Yu、M. Salameh、D. Niu 和 S. Jui，《用于关键词检测的神经架构搜索》，发表于国际语音通信协会年会 *(Interspeech)*，2020年，第 1982–1986 页。[20] B. U. Pedroni、S. Sheik、H. Mostafa、S. Paul、C. Augustine 和 G. Cauwenberghs，《用于高能效关键词检测的小型脉冲神经网络》，发表于*IEEE*生物医学电路与系统会议 *(BioCAS)*，2018年。[21] S. Myer 和 V. S. Tomar，《利用时延神经网络进行高效关键词检测》，*arXiv*预印本 *arXiv:1807.04353*，2018年。[22] V. Garg、W. Chang、S. Sigtia、S. Adya、P. Simha、P. Dighe 和 C. Dhir，《用于硬件高效语音触发检测和误触发抑制的流式Transformer》，发表于国际语音通信协会年会 *(Interspeech)*，2021年，第 4209–4213 页。[23] Y. Wang、H. Lv、D. Povey、L. Xie 和 S. Khudanpur，《采用无对齐无格MMI的唤醒词检测》，发表于国际语音通信协会年会 *(Interspeech)*，2020年，第 4258–4262 页。[24] T.-Y. Lin、P. Goyal、R. Girshick、K. He 和 P. Doll´ar，《用于密集目标检测的焦点损失》，发表于*IEEE*国际计算机视觉会议 *(ICCV)*，2017年，第 2980–2988 页。

[25] Laurens Van der Maaten 和 Geoffrey Hinton，《使用 t-SNE 可视化数据》，机器学习研究期刊，第 9 卷，第 11 期，2008 年。

[26] P. Warden，《Speech Commands：用于有限词汇语音识别的数据集》，*arXiv* 预印本 *arXiv:1804.03209*，2018年。[27] J. Kennedy、S. Lemaignan、C. Montassier、P. Lavalade、B. Irfan、F. Papadopoulos、E. Senft 和 T. Belpaeme，《人机交互中的儿童语音识别：评估与建议》，载于*ACM/IEEE* 人机交互国际会议（*HRI*）论文集，2017年，第 82–90 页。[28] A. Ephrat、I. Mosseri、O. Lang、T. Dekel、K. Wilson、A. Hassidim、W. T. Freeman 和 M. Rubinstein，《在鸡尾酒会中聆听：用于语音分离的说话人无关音视频模型》，*ACM* 图形学汇刊，第 37 卷，第 4 期，第 109:1–109:11 页，2018年。[29] J. F. Gemmeke、D. P. W. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，《 Audio Set：音频事件本体与人工标注数据集》，载于 *IEEE* 声学、语音与信号处理国际会议（*ICASSP*）论文集，2017年，第 776–780 页。[30] Silero 团队，《Silero VAD：预训练的企业级语音活动检测器（VAD）、数字检测器和语言分类器》，https://github.com/snakers4/silero-vad，2021年。[31] D. Snyder、G. Chen 和 D. Povey，《MUSAN：音乐、语音与噪声语料库》，*arXiv* 预印本 *arXiv:1510.08484*，2015年。[32] L. N. Smith 和 N. Topin，《超级收敛：利用较大学习率极快速地训练神经网络》，载于面向多域行动的人工智能与机器学习应用。SPIE，2019年，第 11006 卷，第 369–386 页。