噪声环境下结合空间先验的端到端方向感知关键词检出

Rui Wang、Zhifei Zhang、Yu Gao*、Xiaofeng Mou、Yi Xu

美的集团人工智能研究中心（上海）有限公司，中国{

wangrui302, zhangzf148, gaoyu11, mouxf, xuyi42}@midea.com

arXiv:2603.09505v1 [eess.AS] 10 Mar 2026

摘要

关键词检出（KWS）对于许多语音驱动的应用至关重要，但在噪声环境下实现稳健的关键词检出仍然面临挑战。传统系统通常依赖单通道输入，并采用将前端增强与关键词检出分开的级联流水线。这种做法无法进行联合优化，从而不可避免地限制了性能。我们提出了一种端到端多通道关键词唤醒框架，利用空间线索提升抗噪能力。空间编码器学习通道间特征，空间嵌入则注入方向先验；融合后的表征由流式骨干网络处理。在多个信噪比（SNR）下进行的模拟噪声条件实验表明，空间建模和方向先验各自都能带来明显优于基线方法的效果，二者结合时表现最佳。这些发现验证了端到端多通道空间建模的有效性，并表明其在复杂声学场景下的目标说话人感知检测中具有巨大潜力。

索引词：关键词检出、端到端、噪声鲁棒性、空间先验

1. 引言

关键词检出（KWS），也称为唤醒词检测（WWD），旨在从连续语音音频流中检测预先设定的唤醒词。关键词检出被广泛用作语音接口的入口，并广泛应用于许多场景，尤其是 Amazon Alexa、Apple Siri、Google Assistant 和 Microsoft Cortana 等商业语音助手 [1] [2]。

语音接口的实际应用（包括关键词识别和自动语音识别）通常需要响应特定目标说话人的指令。同时，真实场景中的声学环境也更加复杂，背景噪声、混响和重叠语音都会降低性能。常见的解决方案是采用级联流水线，在声学模型 [3–9] 前加入前端增强/波束成形阶段。随着麦克风阵列得到更广泛的部署，多通道处理变得愈发重要；研究表明，在具备多通道信号时，利用目标声源的空间信息能够显著改善分离/提取效果 [10–12]。然而，在传统级联系统中，前端与检测器相互分离，妨碍了联合优化，并可能导致目标函数不匹配 [13–15]。即使进行了大量数据增强（例如加入加性噪声、卷积 RIR），在噪声和干扰下保持鲁棒性仍然具有挑战性 [16]。

深度学习的最新进展推动关键词识别转向端到端（E2E）建模，通过联合学习特征提取和检测来完成识别 [17, 18]。WeKws 等工具包提供了标准化流水线和实用的流式处理方案[19]。然而，在嘈杂环境等复杂声学条件下有效识别目标说话人的语音仍然具有挑战性。这些端到端方法仍存在两个主要缺陷。首先，许多端到端系统假设输入为单通道，或仅通过简单的通道堆叠处理多通道信号，未能充分利用空间线索，也缺少可学习的空间模块 [14]。其次，系统很少显式建模方向信息，因此在多说话人条件下，对目标说话人的感知能力有限。因此，在实际部署中，噪声鲁棒性和干扰抑制能力仍受到限制。要弥补这些不足，需要一种端到端框架，在识别器中融入空间建模，并支持注入方向感知先验。

本文提出了一种适用于噪声环境下多通道关键词唤醒的端到端方向感知框架。其核心是端到端集成以下模块：（i）用于特征提取的空间编码器；（ii）利用空间线索注入方向感知先验信息的空间嵌入；以及（iii）流式关键词唤醒模型。空间编码器从多麦克风信号中学习通道间特征，而空间先验则与多通道特征相关联，从而增强噪声环境下针对目标说话人的关键词唤醒能力。与传统的单通道方法以及采用独立前端增强模块的级联系统相比，实验结果表明，所提出的统一架构具有明显优势。本文其余部分安排如下：第 2 节详细介绍所提出的框架，包括架构和流程；第 3 节介绍实验评估；第 4 节总结全文。

2. 端到端方向感知关键词识别

2.1. 概述

端到端关键词识别系统通常包含三个组成部分：(i) 作为预处理的特征提取模块（例如波形或频谱特征），(ii) 将帧/片段映射为高层表示的声学编码器（主干网络），以及 (iii) 将后验轨迹转换为关键词触发信号的分类器。大多数现有的端到端流程都是单通道的，并使用无对齐池化目标或序列准则（例如 CTC），从连续语音中学习检测预定义命令 [20] [21], ，但通常不会显式建模空间线索。

*通讯作者。

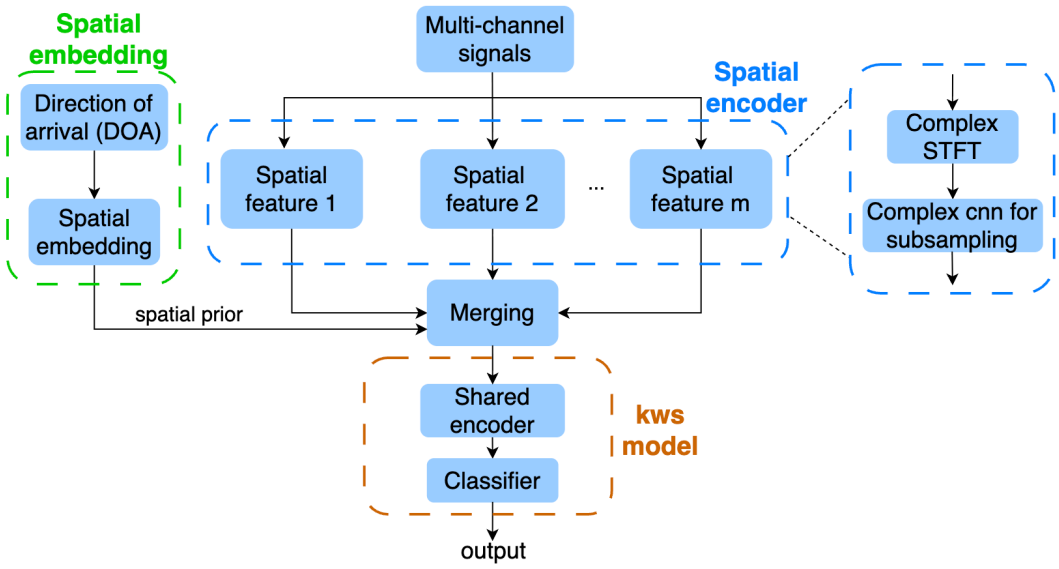


图 1：所提出的关键词识别框架结构。

2.2. 提出的端到端空间感知框架

为解决上述局限，我们提出一种关键词识别框架，将端到端方向感知集成作为整体，联合学习：(i) 从多通道信号中提取空间信息的空间编码器，(ii) 注入空间先验信息的空间嵌入，以及 (iii) 流式关键词唤醒模型。前两者提供互补的条件化信息——基于音频的空间声学特征和基于标签的方向线索——同时，整个网络在标准损失函数下围绕最终的关键词识别目标进行优化。数据流经整体框架的示意图如图 1 所示。本节接下来将详细介绍所提框架的各个模块，以及训练和推理流程。

2.2.1. 数据准备与输入呈现

在我们的实现中，我们使用多通道系统。多通道波形被转换为时频（TF）域中的复数频谱特征，记为 $X_m(f,t)$，其中 $m =$ 表示通道数。这些复数频谱特征保留了通道间的相位和幅度关系。因此，它们包含可供下游模块使用的空间线索。后续章节将详细介绍数据生成和特征配置。

2.2.2. 空间编码器

我们直接对多通道复数频谱特征 $X_m(f,t)$进行处理，采用两阶段二维卷积下采样器：先进行复数二维卷积（沿时间—频率方向设置步幅），再接 ReLU，然后使用轻量级实数二维卷积并设置步幅。这样可将时间分辨率降低 ($e.g., T \to T'$)。空间编码器 $\text{Encoder}_{\text{SP}}$ 输出时间对齐的特征序列。

$$H \in \mathbb{R}^{B\times T'\times d} = \text{Encoder}_{\text{sp}}(X_m(f,t)) \qquad (1)$$

其中 B 表示批量大小，T' 是下采样后的时间步数，而 d 是特征维度，对应于编码器输出。空间编码器的输出保留通道间线索——类似通道间相位差/电平差（通道间相位差和电平差），无需显式波束合成，并与关键词识别目标进行端到端训练。

2.2.3. 空间嵌入

我们假设在训练和评估期间已知目标信号的到达方向（DOA）；在部署时，同一接口可由任意到达方向估计方法提供输入。我们使用轻量级嵌入网络 Emb 对离散方向标签进行映射，该网络将基于 DOA 的标签 θ 转换为紧凑的先验

$$e_\theta = \text{Emb}(\theta), \quad e_\theta \in \mathbb{R}^d \qquad (2)$$

其中，$\theta \in \{1,\dots,K\}$ 和 K 表示声学空间中的区域索引。DOA 标签 θ 被离散为覆盖阵列视场（FOV）的 K 个角度区域。这里，K控制先验的空间分辨率，近似分箱宽度为 FOV/K。在空间嵌入网络中，使用带 ReLU非线性的两层多层感知机对表示进行细化，随后通过随机失活和层归一化来稳定条件化。我们采用线性特征融合，将空间编码器生成的多通道输出特征与生成的空间先验结合，得到

$$\tilde{H}=H+e_\theta \qquad (3)$$

这种融合使模型偏向目标方向，同时保留声学证据。随后，流式骨干网络会处理融合后的序列 $\tilde{H}$。

2.2.4. 关键词检测模块

我们采用多尺度深度可分离时域卷积（MDTC）骨干网络 [22] 作为共享编码器。MDTC 堆叠了具有不同膨胀率的因果深度时域卷积（DTC）模块，以在流式处理约束下捕获多尺度时序上下文，从而在不增加参数开销的情况下有效扩大感受野。编码器以

˜融合序列 H 作为输入，并为分类器头生成统一表示。推理以流式方式进行，采用因果卷积并缓存左侧上下文；不会使用未来帧。

为支持多个唤醒词，我们在共享编码器后接入相互独立的二元分类器——每个关键词对应一个单节点 Sigmoid 头；所有关键词共用编码器参数。这种设置遵循关键词识别系统中的标准做法 [19], ，既保持骨干网络紧凑，又能对各个关键词独立评分和触发。实际上，这种设计还便于通过仅训练相应的小型头、复用同一个共享编码器来添加或移除关键词。运行时，各个头会输出对应关键词的后验概率；随后通过简单的后验平滑和针对每个关键词设置的阈值，产生稳定的触发信号。

3. 实验评估

我们对所提出的方法和若干基线方法进行了实验评估。本节将介绍实验设置和评估结果。

3.1. 训练与测试设置

3.1.1. 数据集

我们使用 Google Speech Commands v1（GSC v1） [23] 语料库作为训练所用的干净语音和文本标签来源。GSC v1 语料库包含 1,881 位不同说话人录制的 30 个单词，共 64,721 段时长为一秒的录音。我们沿用 [19], 中的训练/验证/测试集划分设置，其中 30 个单词中的 10 个被设为关键词。

3.1.2. 空间信号模拟与到达方向标签

为获得带有到达方向标签的空间化信号，我们使用 gpuRIR [24]合成了多通道录音。[1] 由于 GSC v1 为单通道，每条语音均被视为点声源，并通过与针对特定阵列几何结构生成的房间脉冲响应（RIR）进行卷积来实现空间化。声学环境在所有配置中保持一致：房间的长度在 3–8 m、宽度在 3–5 m、高度在 2.5–4 m 范围内随机采样， RT_{60} 为 0.05–0.8 s，声源与阵列的距离取自 0.5–5 m。

我们模拟了两种麦克风配置。对于双通道设置，采用间距为 3 cm 的线性阵列。目标信号到达方向 θ 被限制在前半球（0°–180°），并离散为六个 30° 空间区域（得到标签 1–6）。对于扩展的三通道设置，空间覆盖范围扩展至完整的 360° 方位角，目标信号到达方向相应地被离散为十二个均匀的 30° 区域（标签 1–12）。在两种配置中，均额外保留标签 0，用于表示基线训练和测试中的“无空间先验”条件。因此，每条渲染后的语音都带有多通道波形和离散区域标签 θ。

3.1.3. 噪声渲染

背景噪声取自包含 17 类噪声的 DEMAND 数据集 [25], 。我们使用其中 12 类进行训练，另外 5 类用于测试。噪声经过空间化处理：与独立采样的多通道 RIR（来自同一模拟器／房间池）进行卷积，然后在训练期间，以均匀采样的信噪比（SNR） $[0, 10]$ dB 与目标信号混合。

3.1.4. 目标函数

训练所使用的损失函数是在 C 个类别（关键词加一个非关键词类别）上计算的交叉熵。将时间轴展平后，我们对有效帧上的损失函数取平均；同时还报告了逐帧准确率，作为训练诊断指标。

3.2. 噪声条件下的关键词识别性能评估

3.2.1. 测试渲染

测试多通道信号使用与训练相同的模拟器、阵列和 DOA 离散化生成，但信噪比固定为 0、5 和 10 dB。对于每个信噪比，我们渲染了一个包含 6835 个测试样本的独立测试集。除非另有说明，我们提出的系统中用于先验信息的空间信息来自所生成 RIR 的渲染元数据。

3.3. 对比系统

3.3.1. 单通道基线

我们使用原始 WeKws 实现作为单通道关键词识别基线，其梅尔滤波器组（Fbank）特征设置与 [19] 中相同。对于空间化的双通道信号，我们仅将通道 1 送入后端，因此未使用任何空间信息。

3.3.2. 增强级联基线

我们在 WeKws 后端之前插入广义旁瓣抵消器（GSC） [26] 波束形成器，以构成级联系统。多通道输入首先使用独立的 GSC 方法进行增强；然后将一路波束输出送入与第 3.3.1 节相同的 WeKws 流程。该基线通过波束形成器保留对空间信息的访问，但将其与识别器解耦。我们用它来评估在相同空间条件下，带有空间增强的级联设计与我们的端到端方法相比表现如何。

3.3.3. 无先验的双通道端到端系统

我们的无空间先验双通道端到端系统（无先验的双通道端到端）：在训练和测试期间，到达方向标签均固定为 $\theta = 0$ （无先验标记）。该基线完全依赖空间编码器内部的卷积层，从两个传感器中隐式提取通道间相位差。其他设置保持不变。

3.3.4. 所提出的双通道空间端到端系统

与第 2.2 节所述相同的带空间先验双通道端到端系统（2通道所提出的空间端到端）。如第 2.2 节所述，到达方向标签（1–6）经过嵌入后与编码器特征融合。

3.3.5. 扩展至三通道系统

为严格评估空间编码器的可扩展性以及更高角分辨率的影响，我们将端到端系统从双通道扩展为三通道系统。此次扩展需要调整输入维度，并由此产生了两种额外配置。

[1] https://github.com/DavidDiazGuerra/gpuRI R

表 1：不同信噪比下关键词识别的平均准确率 *s.*

方法	信噪比 = 0 dB 参数	准确率 (%)
单通道基线 [19]	164k	69.86
增强级联基线	164k	72.19
无先验的双通道端到端	279k	76.92
无先验的三通道端到端	279k	80.23
2通道所提出的空间端到端	**279k**	**77.67**
3通道所提出的空间端到端	**279k**	**79.39**
方法	信噪比 = 5 分贝 参数	准确率(%)
单通道基线 [19]	164k	75.89
增强级联基线	164k	78.19
无先验的双通道端到端	279k	84.26
无先验的三通道端到端	279k	85.85
2通道所提出的空间端到端	**279k**	**85.25**
3通道所提出的空间端到端	**279k**	**85.22**
方法	信噪比 = 10 dB 参数	准确率 (%)
单通道基线 [19]	164k	81.68
增强级联基线	164k	82.87
无先验的双通道端到端	279k	88.57
无先验的三通道端到端	279k	89.40
2通道所提出的空间端到端	**279k**	**89.42**
3通道所提出的空间端到端	**279k**	**89.61**

首先，对于无先验的三通道端到端，声学前端经过调整，可接收三通道堆叠音频输入。在训练和测试阶段，外部到达方向标签均严格固定为 $\theta = 0$（无先验标记）。

其次，对于3通道所提出的空间端到端方法，系统利用额外麦克风提供的更密集空间采样。目标到达方向空间被映射为更精细的12区离散标签（1–12），覆盖完整的 360° 方位角。空间编码器也相应扩展，以适应更大的方向索引词汇表。除输入通道维度和空间嵌入词汇表大小外，其他所有网络超参数和训练准则均与双通道配置保持一致。

4. 结果与讨论

4.1. 整体性能与基线比较

表1总结了结果和模型参数。我们的带有方向先验的双通道端到端关键词检测在所有信噪比下均位居双通道配置之首。在0 dB下，其准确率达到77.67%，相较于单通道WeKws基线相对提升11.18%，较波束形成器级联绝对提升5.48%。

在参数预算相当的情况下，端到端系统的表现始终优于非端到端基线方法。尽管级联系统的表现优于单通道基线方法，但仍受限于前端与后端的分离：增强过程中产生的处理伪影无法与下游识别器联合优化。这证实，在统一的端到端框架中进行联合空间与声学建模，比采用分离式流水线更有效。

4.2. 空间先验的影响

在双通道和三通道配置下，采用空间先验的模型与无先验基线之间的性能差距都相对较小。这主要是因为当前评估中没有明确区分的干扰说话人。缺乏强烈方向性干扰时，仅凭多通道声学特征，在背景噪声下进行稳健的关键词检测就已能获得丰富信息。因此，无先验模型依然表现出色，它们依赖学习到的空间表征，动态适应空间线索，并独立取得优异表现。

尽管如此，方向先验会因空间任务复杂度不同而对模型产生不同影响。在双通道设置中，将 180° 空间离散为6个区域会得到一种粗粒度先验，通常对DOA失配不太敏感。这一宽泛线索在各信噪比条件下，相对于无先验基线都能带来虽小但稳定的增益，即便噪声削弱了声学线索。

相反，三通道系统将先验扩展为包含12个区域的 360° 全方位布局，使空间词汇量翻倍。在较低信噪比（0和5 dB）下，将这种严格的高分辨率先验条件化于被噪声模糊的特征，会增加特征-先验失配的风险。因此，保留数据驱动灵活性的无先验基线略微优于所提出的模型。然而，在10 dB信噪比下，干净的声学特征能够可靠地与复杂的12区域先验对齐。此时，空间先验起到更强的方向偏置作用，使所提出的三通道模型达到总体最高准确率（89.61%）。这表明，先验粒度应与不同噪声强度下空间线索的可靠性相匹配；我们将在更复杂的声学场景中，通过到达方向误差扰动对此作进一步分析。

5. 结论

本文提出了一种端到端多通道关键词识别框架，将可学习空间编码器、方向感知嵌入和流式共享骨干网络融为一体。通过明确利用通道间线索并注入离散空间先验，该系统在相近的参数规模下，比单通道WeKws基线和波束形成增强级联流水线具有更强的噪声鲁棒性。双通道和三通道系统的评估结果表明，即使没有先验信息，该模型仍能保持稳健；而在加入先验信息后，性能还会进一步提升。我们的分析还揭示了一种实用的权衡：在清晰条件下，高分辨率先验可充当精确的空间滤波器，而在强噪声环境中，低复杂度先验则具有更好的容错能力。这些发现为在多种真实声学环境中部署可靠的语音控制接口提供了宝贵的架构指导。

我们的方法具有模块化和可扩展性。在未来的工作中，可以将可训练的DOA估计器集成到多任务框架中，以动态提供分区级或连续的方向信息，从而对空间嵌入进行条件约束。我们还计划在空间编码器旁加入增强前端，构建联合的定位–增强–唤醒流程。预计这种联合优化能够进一步缓解高度混响环境下的特征先验失配。我们将探索更丰富的先验设计，例如概率空间嵌入，以提高系统对DOA失配的鲁棒性。最后，我们将研究端侧的弱监督微调，以便在严格的流式处理延迟和内存约束下部署这一统一系统。

6. 参考文献

[1] M. B. Hoy，《Alexa、Siri、Cortana等：语音助手简介》，《医学参考服务季刊》，第37卷，第1期，第81–88页，2018年。[2] A. H. Michaely、X. Zhang、G. Simko、C. Parada和P. Aleksic，《使用上下文语音识别实现Google Assistant的关键词检出》，载于2017年IEEE自动语音识别与理解研讨会（ASRU）。IEEE，2017年，第272–278页。[3] J. Heymann、L. Drude和R. Haeb-Umbach，《基于神经网络的声学波束成形频谱掩码估计》，载于2016年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2016年，第196–200页。[4] H. Erdogan、J. R. Hershey、S. Watanabe、M. I. Mandel和J. Le Roux，《使用单通道掩码预测网络改进MVDR波束成形》，载于Interspeech，2016年，第1981–1985页。[5] T. Yoshioka、N. Ito、T. Delcroix、A. Ogawa、K. Kinoshita、M. Fujimoto、C. Yu、W. J. Fabian、M. Espi、T. Higuchi等，《NTT CHiME-3系统：面向移动多麦克风设备的语音增强与识别进展》，载于2015年IEEE自动语音识别与理解研讨会（ASRU）。IEEE，2015年，第436–443页。[6] F. Weninger、H. Erdogan、S. Watanabe、E. Vincent、J. Le Roux、J. R. Hershey和B. Schuller，《使用LSTM循环神经网络进行语音增强及其在抗噪自动语音识别中的应用》，载于潜变量分析与信号分离国际会议。Springer，2015年，第91–99页。[7] T. Higuchi、N. Ito、T. Yoshioka和T. Nakatani，《利用时频掩码实现面向噪声环境中在线/离线自动语音识别的鲁棒MVDR波束成形》，载于2016年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2016年，第5210–5214页。[8] D. Bonet、G. C´ambara、F. L´opez、P. G´omez、C. Segura和J. Luque，《语音助手中的唤醒词检测语音增强》，arXiv预印本，arXiv:2101.12732，2021年。[9] C. Yang、Y. M. Saidutta、R. S. Srinivasa、C.-H. Lee、Y. Shen和H. Jin，《通过利用语音增强和语音存在概率提高噪声环境下的鲁棒关键词检出》，载于Interspeech论文集，第2023卷，2023年，第1638–1642页。

[10] L. Parra 和 C. Alvino，“几何源分离：将卷积源分离与几何波束成形相结合”，IEEE 语音与音频处理汇刊，第 10 卷，第 6 期，第 352–362 页，2002 年。

[11] L. Li 和 H. Kawahara，“用于定向语音增强的几何约束独立向量分析”，载于 ICASSP 2020-2020 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2020年，第846–850页。

[12] R. Wang 和 T. Toda，“通过结合全局风格令牌的条件变分自编码器，在噪声欠定条件下进行定向目标说话人提取”，载于 2023年IEEE音频与声学信号处理应用研讨会（WAS-PAA）。IEEE，2023年，第1–5页。

[13] M. L. Seltzer，“弥合差距：使用麦克风阵列实现免提语音识别的统一框架”，载于 2008年免提语音通信与麦克风阵列。IEEE，2008年，第104–107页。

[14] H. Zhang, J. Zhang, 和 Y. Wang，“端-to-端模型与听觉注意力的多通道关键词检测，”arXiv预印本 arXiv:1811.00350，2018。

[15] Q. Yang, Q. Li, 和 Z. Ma，“面向低资源场景关键词检出的深度残差脉冲神经网络”。载于 Interspeech，2022 年，第 3023–3027 页。

[16] T. Higuchi、S. Saxena、M. Souden、T. D. Tran、M. Delfarah和C. Dhir，“基于数据参数的动态课程学习用于噪声鲁棒关键词检出”，载于ICASSP 2021-2021 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2021，第6848–6852页。

[17] K. Zhang、Z. Wu、D. Yuan、J. Luan、J. Jia、H. Meng和B. Song，“用于处理端到端关键词检出中数据不平衡问题的重加权区间损失”，载于INTERSPEECH，第108卷，2020，第2567–2571页。[18] I. L´opez-Espejo、Z.-H. Tan、J. H. Hansen和J. Jensen，“深度口语关键词检出：综述”，IEEE Access，第10卷，第4169–4199页，2021。[19] J. Wang、M. Xu、J. Hou、B. Zhang、X.-L. Zhang、L. Xie和F. Pan，“Wekws：面向生产的轻量级端到端关键词检出工具包”，载于ICASSP 2023-2023 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2023，第1–5页。[20] A. Graves、S. Fern´andez、F. Gomez和J. Schmidhuber，“连接时序分类：利用循环神经网络标注未分段序列数据”，载于第23届国际机器学习会议论文集，2006，第369–376页。[21] S. Jin、Y. Jung、S. Lee、J. Roh、C. Han和H. Cho，“面向流式开放词汇关键词检出的CTC对齐音频-文本嵌入”，arXiv预印本arXiv:2406.07923，2024。[22] J. Hou、L. Xie和S. Zhang，“基于多尺度深度可分离时序卷积的两阶段流式关键词检测与定位”，Neural Networks，第150卷，第28–42页，2022。[23] P. Warden，“Speech Commands：一个用于有限词汇语音识别的数据集”，arXiv预印本arXiv:1804.03209，2018。[24] D. Diaz-Guerra、A. Miguel和J. R. Beltran，“gpurir：一个支持GPU加速的房间脉冲响应仿真Python库”，Multimedia Tools and Applications，第80卷，第4期，第5653–5671页，2021。[25] J. Thiemann、N. Ito和E. Vincent，“Demand：多样化环境中声学噪声的多通道录音集”，（无标题），2013。[26] L. Griffiths和C. Jim，“线性约束自适应波束成形的另一种方法”，IEEE天线与传播汇刊，第30卷，第1期，第27–34页，1982。