arXiv:2602.06271v1 [cs.SD] 6 Feb 2026

使用冻结的预训练卷积神经网络和时间序列模块混合模型检测合成声景中的恐音症触发声音

Kurumi Sashida[1] 和Gouhei Tanaka[1,2]

[1]名古屋工业大学计算机科学系，日本名古屋 466-8555[2]东京大学神经智能国际研究中心，日本东京 113-0033*k.sashida.344@stn.nitech.ac.jp, gtanaka@nitech.ac.jp*

2026年2月9日

摘要

恐音症是一种对特定日常声音（触发声音）耐受度降低的障碍，这些声音可能引发愤怒、恐慌或焦虑等强烈负面情绪反应。这些反应会严重影响日常功能和生活质量。能够选择性检测触发声音的辅助技术有助于减轻痛苦并提升身心健康。本研究探讨声音事件检测（SED），以定位连续环境音频中触发声音出现的时间区间，为此类辅助支持奠定基础。鉴于现实世界中的恐音症数据稀缺，我们利用音频合成技术生成专为恐音症触发声音检测设计的合成声景。随后，我们使用基于卷积神经网络的混合模型执行触发声音检测任务。这些模型结合了冻结的预训练 CNN 主干网络进行特征提取，以及可训练的时间序列模块，例如门控循环单元（GRU）、长短期记忆网络（LSTM）、回声状态网络（ESN）及其双向变体。我们采用常见的声音事件检测指标评估检测性能，包括多声部声音检测得分 1（PSDS1）。在多类别触发声音事件检测任务中，双向时序建模始终能提升检测性能，其中双向GRU（BiGRU）的整体准确率最高。值得注意的是，双向ESN（BiESN）仅需优化读出层即可达到具有竞争力的性能，同时所需的可训练参数数量少几个数量级。我们还通过少样本“进食声”检测任务模拟用户个性化，该任务最多使用五个支持片段，并对比BiGRU和BiESN。在这一严格的适应设置下，BiESN表现稳健且稳定，表明轻量级时序模块有望用于个性化恐音症触发声音事件检测。

关键词： 辅助技术、音频机器学习、回声状态网络、个性化、声音事件检测

1 引言

恐音症是一种对特定声音或相关线索（诱发因素）的耐受性降低的障碍 [1]。诱发因素通常具有重复性，最常见的是口面部声音（例如进食、吸鼻子或呼吸），但也可能包括键盘打字等声音 [1,2]。接触触发声音会引发强烈的负面情绪（例如愤怒、焦虑或恐慌），并伴随自主神经唤醒，有时被描述为一种不由自主的战或逃反应 [3]。恐音症反应

往往取决于情境，并受到声源身份的调节。例如，人们通常会对他人发出的声音产生反应，而自己发出相同的声音时则能够忍受。这些反应并非仅由响度决定 [1]。恐音症会损害学业和职业功能，并导致人际关系恶化，从而限制社会参与、降低生活质量 [1]。为了应对日常生活中的这些诱发因素，人们会采取一些措施，例如离开现场、使用耳塞或降噪耳机减少听觉输入，或用音乐、白噪声等其他声音加以掩蔽 [2]。尽管这些方法可能带来即时缓解，但持续回避可能会提高唤醒水平并增强对诱发因素的觉察，从而导致症状持续存在 [4]。此外，针对相关声音耐受不良状况的研究证据表明，持续减少听觉输入（例如过度使用听力保护设备）可能改变响度感知，这与中枢听觉增益的适应性变化相一致 [5,6]。不应假定恐音症患者会对触发声音产生习惯化，因为症状可能持续存在，在某些情况下还可能随时间加重 [4]。因此，需要能够有选择地检测触发声音，并在必要时对其进行衰减或抑制，同时保留对周围声学环境的感知的辅助技术。要实现这种有选择的实时抑制，需要一种能够识别触发声音起始时间、结束时间和类别的检测机制。

先前关于恐音症的研究主要刻画了对预先选定的听觉刺激以及精心整理的刺激集的反应 [7,8]。近年来，基于学习的方法也被用于识别或分类与恐音症相关的声音 [9–11]。然而，音频分类通常是片段级任务，针对时长从数秒到十余秒的声音片段预测一个类别（或多个类别）；它并不能直接识别连续音频中的具体发生区间（即起止时间）（图1）。在连续的日常音频中检测触发声音何时出现以及出现何种触发声音，目前仍相对缺乏研究。

因此，我们研究声音事件检测（SED），以同时估计触发声音类别及其时间发生位置，作为面向恐音症的辅助应用的基础。特别是，为了在未来实现个性化和实时的恐音症缓解技术，必须开发能够在有限计算资源下进行设备端运行的高效声音事件检测模型。声音事件检测旨在识别发生了哪些声音事件，并在录音中定位其起止时间 [12]。在以声学场景与事件检测和分类（DCASE）挑战赛任务4为中心的声音事件检测研究中，基准测试通常用大规模包含强标注的合成声景来补充真实录音，以应对带有强标签的真实世界数据稀缺的问题 [13]。具体而言，家庭环境声音事件检测（DESED）数据集既包括来自AudioSet的真实家庭录音 [14] ，也包括使用Scaper生成的合成片段 [15], ，这些片段具有可用于监督训练的精确事件时间戳 [13]。作为一种模型框架，卷积循环神经网络（CRNN）常被用于DCASE挑战赛 [13]。CRNN将用于时频模式提取的卷积神经网络（CNN）与用于时间上下文建模的循环神经网络（RNN）及其变体相结合。

值得注意的是，恐音症触发音种类繁多，且因人而异，因此很难收集到能够充分涵盖目标触发类别的强标注真实世界录音 [1, 16]。

本研究遵循 DESED/DCASE Task 4 [13, 17], 的合成流程，使用 Scaper [15] 生成强标注的恐音症触发声景，并以监督方式训练多类别声音事件检测模型。为确保模型架构轻量，我们冻结预训练的 CNN 主干网络，并训练序列模块。我们评估了彼此相近的口面类别（例如咀嚼、吸鼻子和呼吸）的判别性能与事件边界的时间定位能力，以及咳嗽、清嗓、打字和时钟滴答声等短暂脉冲事件的相关表现。

我们每 在恐音症触发声音事件检测任务上进行数值实验，采用一种混合模型 del

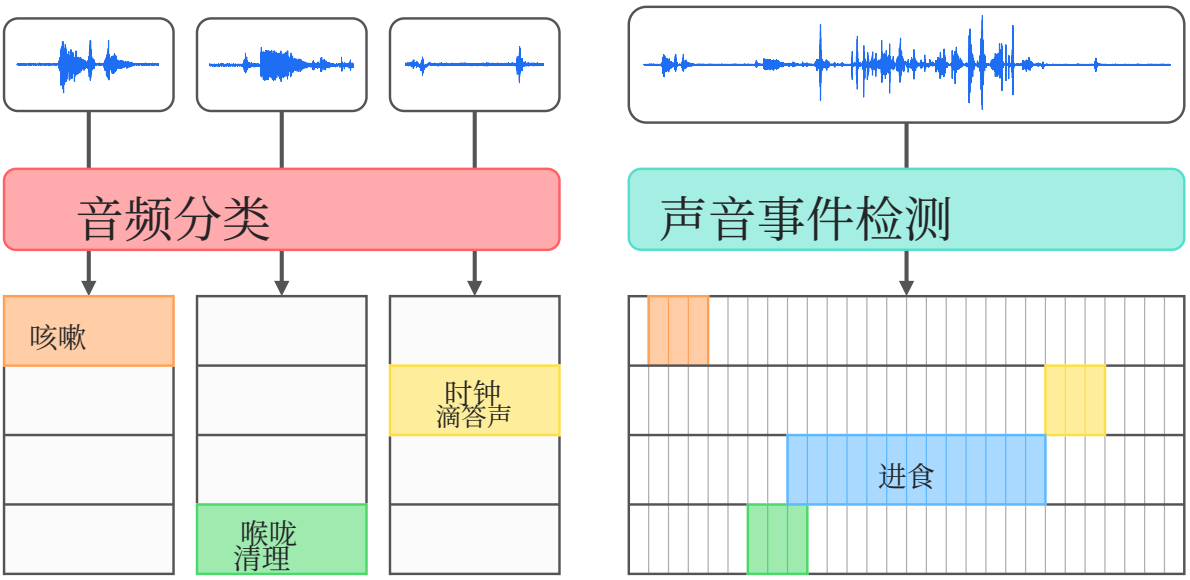


图1：音频分类与声音事件检测（SED）概览。左：音频分类为音频片段分配片段级标签，但不进行时间定位。右：声音事件检测旨在识别事件类别，并估计其在音频片段中的起止时间。

该模型结合了冻结的预训练CNN特征提取器和可训练时间序列模块。作为时序模块，我们采用门控循环单元（GRU）、长短期记忆网络（LSTM）、回声状态网络（ESN）及其双向变体。首先，我们在包含七种触发类别的合成声景上执行多类别触发声音事件检测任务。结果表明，双向建模能显著提升性能，其中双向门控循环单元（BiGRU）的准确率最高，而双向ESN（BiESN）则在性能与参数效率之间实现了颇具吸引力的平衡。其次，为应对触发因素的个体差异，我们模拟了一个聚焦于“进食声”类别的少样本用户自适应触发声音事件检测任务。在这一数据稀缺场景中，支持片段最多只有五个，与完全可训练的BiGRU相比，BiESN表现出更强的稳定性，显示出其在有限数据场景下实现稳健个性化的潜力。

2 数据集合成

目前，公开可用的标准化恐音症触发声音检测数据集仍然稀缺。我们构建了一个用于监督式多类声音事件检测的强标注合成声景数据集，涵盖七种厌恶声音相关触发类别：五种人类产生的口鼻腔/喉部声音（进食、吸鼻子、清嗓、咳嗽和呼吸声），以及两种重复性非口鼻腔声音（作为人类产生的机械声音的打字声，以及作为机器产生的机械声音的时钟滴答声）。这些类别的选择依据是以往关于常见触发声音的报告和实际数据可用性 [7,16]。遵循 DESED/DCASE任务 4 的方案，我们使用 Scaper声景合成库，将前景事件混入背景录音中，合成时长为 10 秒的声景。强标签（起始/结束时间）直接从合成元数据中获得 [13,15]。

前景事件候选音频取自以下公开语料库：自由开放恐音症刺激（FOAMS）数据库 [8], 厌恶声音视听触发档案（MATA） [18], Freesound 50K数据集（FSD50K） [19], 环境声音分类数据集（ESC-50） [20], 和 VocalSound [21], ；背景录音则取自 Sound in the Noise数据库 [22], 坦佩雷理工大学声学场景

表 1：合成恐音症数据集的统计信息。源文件表示用于合成的唯一孤立录音数量，合成片段表示包含该类别的生成声景数量，总时长表示每个类别所有事件实例的累计时长。

类别	源文件 合成文件	总计 总时长 片段 事件		(分钟)
呼吸	353	2962	3706	106.9
咳嗽	6020	3536	5111	66.1
进食	842	2468	3431	76.1
吸鼻子	4111	3674	5667	60.5
清嗓	3480	3376	4897	62.8
打字	630	2141	2746	87.2
时钟滴答声	168	1848	2008	138.8
总计	15604	10000	27566	598.3

2017 年数据集 [23], 以及音乐、语音和噪声语料库 [24], ，并排除了可能包含任何目标类别的片段。

由于源片段可能包含其他声音（例如语音或操作噪声），我们分两个阶段筛选前景样本：(i) 使用 Yet another Audio MobileNet Network (YAM-Net) [25] 筛选各目标类别的候选片段；(ii) 手动审查 YAMNet 未能分类为目标类别的片段，除非其中明显能听到独立且非目标的事件，否则予以保留。表 1 列出了每个数据集各类别筛选后的源片段数量。合成时，我们从较长的录音中随机截取事件片段，并将移调幅度均匀设定在 -3 至 $+3$ 半音之间。仅对时钟滴答声类别，我们使用谱门控降噪（noisereduce）来减轻滴答声之间持续存在的嘶嘶声。合成前，我们按 3:1:1 的比例将筛选后的源片段划分为训练集/验证集/测试集，并确保同一源录音不会出现在多个划分中。随后，我们从各划分中合成时长为 10 秒的声景，分别为训练集、验证集和测试集生成 6000/2000/2000 个片段。

3 方法

3.1 概述：混合模型

图 2 总结了本研究所采用混合模型的端到端流程。我们冻结预训练的 CNN 主干网络，仅训练时序模块和共享线性读出层，以便在固定前端的条件下隔离时间建模的影响。为系统评估时间建模对多声部声音事件检测的影响，我们采用统一视角考察 RNN 变体。虽然 CRNN 通常将卷积前端和循环后端视为一个整体进行训练 [12], ，但我们的目标是在资源受限的条件下隔离时间动态的贡献。

我们在广义门控状态动态的统一框架下，对时序模块（GRU、LSTM、ESN）进行形式化描述，同时采用线性层作为非时序基线。设 $\mathbf{z}_t$ 为由卷积神经网络在时间 t 提取的输入特征，$\mathbf{s}_t$ 为时序模块的内部状态。我们将一般更新规则定义为保留过去上下文与整合新信息之间的平衡，具体如下：

$$\text{Update:} \quad \mathbf{s}_t = \mathbf{\Gamma}_r \odot \mathbf{s}_{t-1} + \mathbf{\Gamma}_u \odot \tilde{\mathbf{s}}(\mathbf{z}_t, \mathbf{h}_{t-1}), \tag{1}$$

$$\text{Output:} \quad \mathbf{h}_t = \mathbf{\Gamma}_o \odot \phi(\mathbf{s}_t), \tag{2}$$

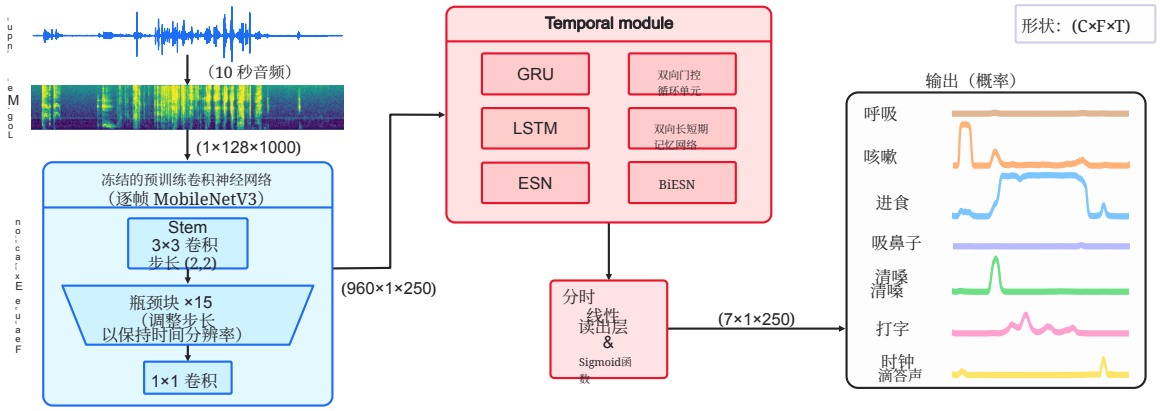


图2：冻结的 CNN + 时序模块流水线，用于逐帧多标签声音事件检测。预训练的逐帧 MobileNetV3 骨干网络提取嵌入$\mathbf{z}_t$，这些嵌入由时序模块（GRU/LSTM/ESN；单向或双向变体）处理，再通过按时间共享的线性 +Sigmoid函数读出层，得到逐帧类别后验概率。

其中，$\mathbf{h}_t$ 为显式呈现的隐藏状态，$\tilde{\mathbf{s}}$ 为由输入和前一隐藏状态导出的候选状态，ϕ 为激活函数。向量 $\mathbf{\Gamma}_r$、$\mathbf{\Gamma}_u$ 和 $\mathbf{\Gamma}_o$ 分别表示保留因子、更新因子和输出 因子，$\odot$ 表示逐元素乘法。

所有模块通过共享的线性读出层，为 C 类别输出逐帧多标签后验概率：$\hat{\mathbf{y}}_t = \sigma(\mathbf{W}_o\mathbf{h}_t + \mathbf{b}_o)$，其中$\mathbf{W}_o \in \mathbb{R}^{C\times H}$和$\mathbf{b}_o \in \mathbb{R}^C$是可学习的读出层参数，$\mathbf{h}_t$ 表示时序模块生成的隐藏状态向量。如第3.3节所述，我们根据各方法对这些门控因子的约束方式（固定标量或自适应门）以及参数的可训练性对其进行分类。

3.2 冻结的预训练 CNN 主干网络

我们使用 EfficientSED [26] 中的逐帧 MobileNet（fmn10）作为冻结前端。该骨干网络由 MobileNetV3 衍生而来，通过移除全局池化并调整步长来保留时间分辨率。给定一个输入音频片段，它会生成一系列嵌入$\{\mathbf{z}_t\}_{t=1}^{T}$，时间分辨率为 40 毫秒。对于 10 秒的片段，$T = 250$。我们使用预训练权重初始化该骨干网络，并保持其冻结。

3.3 模型实例

我们评估了统一框架的四种实例，涵盖从无记忆基线到复杂门控架构的不同类型。对于 RNN 变体（GRU、LSTM、ESN），我们采用声音事件检测模型，并分别设置为单向和双向。双向形式遵循标准方法，即分别沿正向和反向时间方向运行动态过程（式 (1)），并拼接得到的状态$\mathbf{h}_t = [\mathbf{h}_t^{\text{fwd}}\,;\,\mathbf{h}_t^{\text{bwd}}]$ [27]。

3.3.1 线性基线

线性模型作为非时序基线，表示不保留任何历史信息的情况。在统一框架中，这可视为一种退化情形，其中

- $\mathbf{\Gamma}_r = \mathbf{0}$（不保留），
- $\mathbf{\Gamma}_u = \mathbf{1}$，$\mathbf{\Gamma}_o = \mathbf{1}$。

内部状态实际上就是瞬时输入投影 $\mathbf{s}_t = \tilde{\mathbf{s}}(\mathbf{z}_t)$。共享分类器独立应用于每一帧。

3.3.2 门控循环单元

GRU [28] 引入了自适应门控，其中保留因子和更新因子相互耦合，并从数据中学习。在统一框架中，这对应于如下集合：

- $\mathbf{\Gamma}_u = \mathbf{u}_t$ （由神经网络计算的更新门），
- $\mathbf{\Gamma}_r = 1 - \mathbf{u}_t$ （耦合保留），
- $\mathbf{\Gamma}_o = \mathbf{1}$。

这种耦合迫使模型在遗忘过去的知识和写入新信息之间取得平衡。我们采用包含两层、每个方向隐藏维度为 $H = 256$ 的堆叠门控循环单元，这种配置常见于标准 CRNN 基准模型 [29]。在层间以及最终分类头之前使用丢弃率为 $p = 0.3$ 的丢弃，以防止过拟合。我们选择门控循环单元作为主要基准测试，因为它代表了当前声音事件检测中时序聚合的最先进标准。

3.3.3 长短期记忆

长短期记忆网络 [30] 在该框架中具有最灵活的动态特性，其门彼此完全独立，并将记忆单元 $\mathbf{s}_t$ 与隐藏输出 $\mathbf{h}_t$分离。在统一框架中，这相当于以下设置：

- $\mathbf{\Gamma}_r = \mathbf{f}_t$ （遗忘门，独立），
- $\mathbf{\Gamma}_u = \mathbf{i}_t$ （输入门，独立），
- $\mathbf{\Gamma}_o = \mathbf{o}_t$ （输出门）。

这种解耦使长短期记忆网络能够在较长时间内保留记忆，而不必在每个时间步都将其传递至输出。与门控循环单元类似，我们采用包含 $H = 256$ 个单元和 $p = 0.3$ 丢弃的两层架构。我们将长短期记忆网络作为高复杂度基线，以确定其显式记忆管理和更大的参数空间是否能相较于更精简的门控循环单元和回声状态网络架构带来显著的性能提升，或者对于此任务而言，额外的计算成本是否并无必要。

3.3.4 回声状态网络

为了突破仅冻结卷积神经网络特征提取器所能达到的参数效率极限，我们在储备池计算框架中采用 ESN，作为完全可训练的 RNN 变体的轻量级替代方案。ESN 是 RNN 的一种特殊形式，其循环动态是固定的，而非通过学习获得 [31]。更一般的模型是泄漏积分型回声状态网络，它可以灵活调整储备池时间尺度 [32]。在统一框架中，这可由以下恒定标量门控因子表示：

- $\mathbf{\Gamma}_r = 1 - \alpha,$
- $\mathbf{\Gamma}_u = \alpha, \quad \mathbf{\Gamma}_o = \mathbf{1}.$

其中，$\alpha \in (0, 1]$ 是一个称为泄漏率的超参数。关键在于，与 GRU/LSTM 不同，定义 $\tilde{\mathbf{s}}$ 的循环权重是随机初始化并冻结的。只有最终的读出权重 $\mathbf{W}_o$ 需要训练。这种设置用于检验固定随机投影在时间上下文建模中的效果。

为了充分发挥储备池计算方法的潜力，我们使用 Optuna 框架 [33] 优化回声状态网络的超参数以及读出层的训练配置。

我们采用树结构帕岑估计器（TPE）采样器，针对每种训练集规模在 150 次试验中高效探索配置空间。在多类别触发声音事件检测设置中，我们使用完整训练集（6000 个片段）开展一项研究。在少样本用户自适应设置中，我们针对每种支持集规模 $K \in \{1, 2, 3, 4, 5\}$ 分别开展研究，其中 K 表示支持片段的数量。优化目标是最大化验证集上的 PSDS1 分数。回声状态网络动态的搜索空间包括谱半径 ρ（在 0.1 到 1.8 之间均匀采样）、泄漏率 α （在 0.05 到 1.0 之间均匀采样）以及输入缩放因子 σ_{in} （在 0.01 到 5.0 之间对数均匀采样）。对于监督读出层，我们将学习率 η 在 10^{-4} 到 3×10^{-3} 的对数均匀范围内进行调优。其他结构参数，如储备池密度（$d = 0.1$）和随机稀疏拓扑，则保持固定，以维持一致的储备池复杂度。对于双向回声状态网络变体，超参数在正向储备池与反向储备池之间共享，以简化搜索过程。

在多类别触发声音事件检测任务中，我们将储备池节点数设为 1,024，以提供足够的容量来区分多个重叠类别（见第 4 节）。对于少样本用户自适应触发声事件检测，我们缩小储备池规模，使其与 GRU/LSTM 的容量相匹配，从而减轻支持数据有限时的过拟合问题（第 4 节）。

3.4 评估指标

我们使用宏平均 F1 分数（包括基于片段和基于事件的 F1 分数）以及多声道声音检测分数（PSDS） [34]，评估模型在厌声症触发音频事件检测任务（见第 4.1 节）上的表现。对于基于片段的 F1 分数（$F1_{\text{seg}}$），我们按逐秒片段评估预测结果。基于事件的 F1 分数（$F1_{\text{event}}$）是衡量区分能力和精细定位的严格指标：只有当类别正确，且起始/结束时间与标注相差不超过 200 毫秒时，才计为真正例。遵循 DCASE任务4 的评估协议，PSDS 通常在两种应用场景下报告，分别记为 PSDS1 和 PSDS2 [35]。在本研究中，为评估不依赖固定决策阈值的检测鲁棒性，我们报告 PSDS1（场景 1）。由于我们的应用需要对短暂、瞬态的触发事件进行高时间精度检测，因此我们选择场景 1（$\rho_{\text{DTC}} = 0.7$、 $\rho_{\text{GTC}} = 0.7$ 和 $\alpha_{\text{CT}} = 0$），而非场景 2。严格的 0.7 重叠阈值可确保检测事件在时间上得到精确定位，而 $F1_{\text{event}}$ 严格的分类约束则能有效监测跨触发混淆。所有指标均归一化至 $[0, 1]$。

3.5 优化与训练详情

如第3.2节所强调，CNN 主干网络保持冻结；梯度更新仅应用于第3.3节所述的时序模块。对于可训练模块（线性/门控循环单元/长短期记忆网络），我们使用 Adam [36]，以 $\eta = 1 \times 10^{-4}$的学习率和 1024 的批次大小对所有模块参数进行端到端优化。对于 ESN 模块，仅更新线性读出参数；读出层的学习率由 Optuna（第3.3.4节）选定，批次大小保持一致，以确保公平比较。

我们使用验证集选择所有推理时超参数。首先，选择验证集 PSDS1（场景 1）表现最佳的训练检查点。其次，在验证集上调整中值滤波窗口大小，使 PSDS1 最大化，以抑制短暂的虚假激活并减少过度分割。最后，在报告 F1 时，我们通过最大化基于事件的 F1，在验证集上确定类别级决策阈值 $F1_{\text{event}}$，并使用相同的阈值计算$F1_{\text{seg}}$ 。所有模型均使用 PyTorch [37]实现。

4 实验

我们在两种互补的设置下评估所提出的方法：(i) 在合成声景上进行多类别触发声音事件检测；(ii) 针对恐音症个性化场景进行建模的少样本用户自适应触发声事件检测

下文中，我们将基于卷积神经网络的混合模型简称为其后续时序模块的名称。

4.1 任务与设置

4.1.1 多类别触发声音事件检测

在此任务中，我们评估混合模型在七类触发声音事件检测问题上的性能。我们采用第 2 节构建的合成数据集，并使用固定的训练集/验证集/测试集划分（6000/2000/2000 个片段），以确保结果可复现。主要目标是在第 3 节所述的统一框架下，评估不同的时序模块区分和定位口鼻与机械类触发声音的能力。具体而言，我们将 ESN、GRU 和 LSTM 的单向与双向变体与线性基线进行比较。除了检测准确率外，我们还分析各模型的可训练参数数量，以评估性能与模型复杂度之间的权衡。评估采用第 3.4 节定义的指标，重点关注 PSDS1，以衡量时间定位检测性能，同时考察 $F1_{\text{event}}$ 和 $F1_{\text{seg}}$。

4.1.2 少样本用户自适应触发声音事件检测

为模拟用户针对特定触发声配置系统的个性化场景，我们针对“进食声”类别开展了少样本检测实验。我们选取包含明显咀嚼声或咂嘴声的片段，以代表特定用户的触发声特征。我们采用受 DCASE Task 5 [38], 启发的支持集-查询集协议，并改变支持片段数量（样本数），其中 $K \in \{1, 2, 3, 4, 5\}$。我们特别比较了双向门控循环单元和 BiESN，作为少样本个性化的代表性模型。对于每个 K，我们仅使用 K 提供的支持片段对双向门控循环单元和 BiESN 模型（定义见第 3 节）进行适配，并将这些片段中的非目标片段用作负样本。虽然我们采用了与多类别任务相同的模型架构，但将 BiESN 的储备池规模设为 $N = 256$，因为初步实验表明，该容量足以满足二分类检测任务的需求。我们定义了一个固定的支持池，其中包含八段经过筛选的进食声录音。每次运行时，我们使用不同的随机种子从该支持池中抽取 K 个支持片段，因此每次运行的支持组合各不相同。我们报告每个 K在 10 次运行（使用不同随机种子）中的均值和标准差。

4.2 结果

4.2.1 多类别触发声音事件检测

表 2 总结了七类别触发音频事件检测任务的检测性能及可训练参数数量。双向时序建模能够持续提升各时序模块的性能：GRU 的 PSDS1 从 0.55 提高到 0.63，LSTM 从 0.52 提高到 0.60，ESN 从 0.49 提高到 0.55，这表明在每段 10 秒音频片段中利用过去与未来上下文是有益的。在所有方法中，双向门控循环单元的整体表现最佳（PSDS1 =0。63， $F1_{\text{event}} = 0.65$， $F1_{\text{seg}} = 0.85$），双向长短期记忆网络紧随其后（PSDS1 $= 0.60$， $F1_{\text{event}}$ =0。64， $F1_{\text{seg}} = 0.84$）。

尽管可训练参数量显著更少，BiESN 仍能达到具有竞争力的准确率。具体而言，BiESN 的 PSDS1 $= 0.55$ ($F1_{\text{event}} = 0.57$) 为 $F1_{\text{event}} = 0.56$，表现与单向 GRU （ PSDS1 $= 0.55$， $F1_{\text{event}} = 0.56$）相当，而仅需 14,343 个可训练参数，约比 GRU（1,037,319 个参数）少 $72\times$。总体而言，在训练成本和模型大小受限时，BiESN 在检测能力与模型复杂度之间实现了较好的权衡，但与完全可训练的时序模块相比，其峰值性能有所下降。

表2：多类别触发声音事件检测（七个类别）的性能。

模型	方向	PSDS1	事件 F1	分段 F1	可训练参数
线性	–	0.47	0.52	0.80	6,727
GRU	Uni	0.55	0.56	0.81	1,037,319
	Bi	0.63	0.65	0.85	2,221,831
LSTM	Uni	0.52	0.55	0.81	1,300,487
	Bi	0.60	0.64	0.84	2,879,239
ESN	Uni	0.49	0.52	0.78	7,175
	Bi	0.55	0.57	0.81	14,343

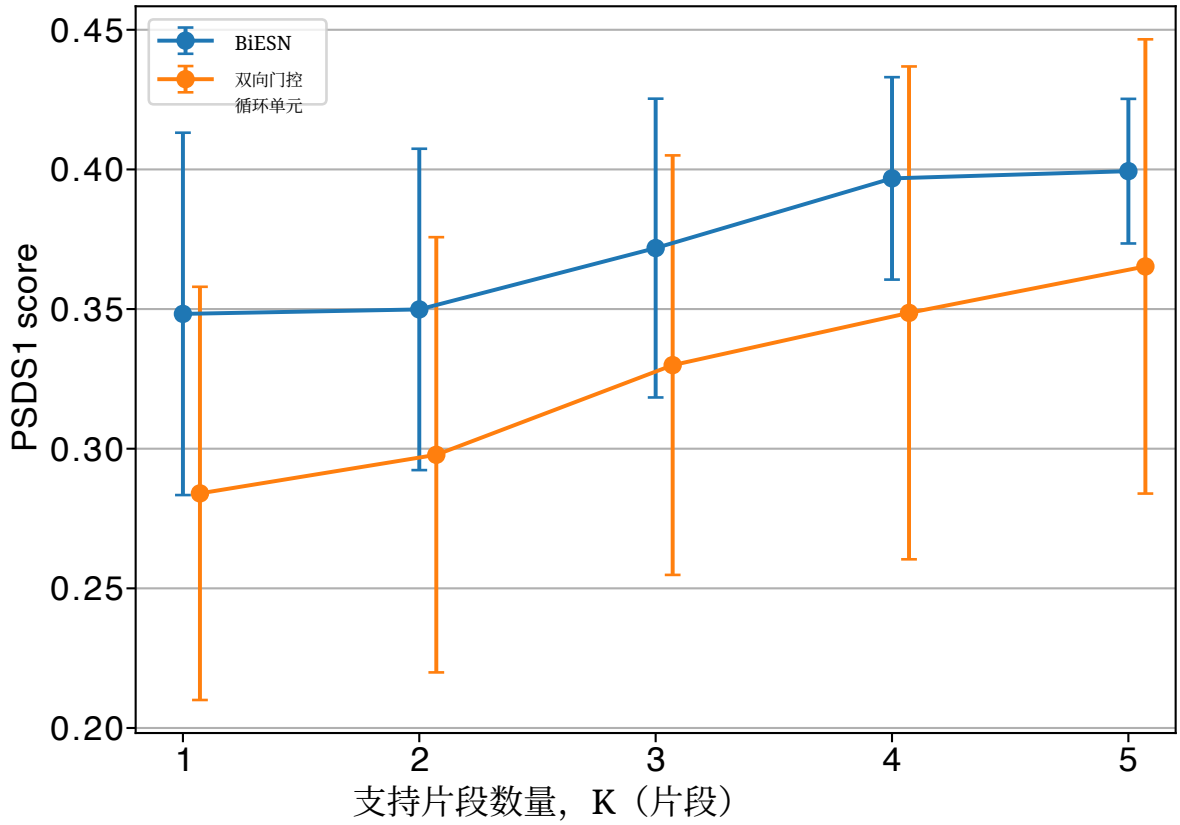


图3：少样本用户自适应触发声事件检测性能，以 PSDS1 随支持片段数量的变化进行衡量 K。在每次运行中，使用不同的随机种子从包含八段录音的固定池中采样 K 支持片段。各点表示 10 次运行的均值，误差线表示 ± 1 标准差。

4.2.2 少样本用户自适应触发声事件检测

图 3 绘出了用户自适应“进食声”检测任务中 PSDS1 随支持片段数量变化的情况 K。在此设置下，BiESN 在所有 K 中均取得了持续高于双向门控循环单元的平均 PSDS1；而双向门控循环单元在不同运行中的波动更大，这表明在数据有限时，其表现更容易受到具体支持集选择和训练随机性的影响。总体而言，结果表明，在我们受限的个性化设置下，采用固定储备池进行仅读出层自适应，可获得更可靠的少样本性能。

5 结论

本研究探讨了恐音症触发声音事件检测，这是迈向端侧应用、缓解恐音症不良影响的基础一步。针对该领域缺乏标准化数据集的问题，我们通过合成声景整理并构建了一个强标注数据集，这是本研究的主要贡献之一。这种方法能够大规模生成精确的起始/结束时间标注，克服了为细微的口腔/鼻腔触发声收集大型真实世界数据集的难题。基于该数据集，我们评估了基于卷积神经网络的混合声音事件检测模型，并在多类别和少样本用户自适应设置下开展实验，以探索切实可行的个性化方案。研究发现，模型架构存在一个关键权衡：虽然双向门控循环单元取得了最高的检测准确率，但 BiESN 以少几个数量级的可训练参数和更简单的仅读出层训练，实现了具有竞争力的性能。这一发现对于未来个性化、实时的恐音症缓解技术尤为重要。

尽管取得了这些令人鼓舞的结果，将合成数据训练与真实世界部署衔接起来仍面临挑战。由于当前性能依赖于合成声景，因此必须验证模型在真实环境多样声学条件下的鲁棒性。未来工作将通过领域自适应解决合成到真实场景的领域差距，并评估流式推理的系统延迟。此外，鉴于恐音症具有高度个体化的特点，我们计划将该框架拓展至用户自适应检测（少样本个性化）和细粒度类内建模。例如，使系统能够在更广泛的“进食”类别中区分“吸食声”和“咀嚼脆响”等特定声学模式至关重要，因为某种声音可能会触发某位用户的症状，而另一种则不会。这种方法将使我们更接近于提供契合个体需求、适用于日常生活的实际缓解方案。

参考文献

[1] S. E. Swedo、D. M. Baguley、D. Denys、L. J. Dixon、M. Erfanian、A. Fioretti、P. J. Jastreboff、S. Kumar、M. Z. Rosenthal、R. Rouw 等，“恐音症的共识定义：一项德尔菲研究”，神经科学前沿，第16卷，第841816页，2022年。[2] Y. Ozuer、R. Cima、E. Kestens和I. Van Diest，“理解恐音症体验：一项混合方法研究”，心理学前沿，第16卷，第1493676页，2025年。[3] M. Edelstein、D. Brang、R. Rouw和V. S. Ramachandran，“恐音症：生理学研究与病例描述”，人类神经科学前沿，第7卷，第296页，2013年。[4] T. Dozier和N. Mitchell，“理解恐音症本质的新型五阶段模型：一种条件性厌恶反射障碍”，*F1000Research*，第12卷，第808页，2023年。[5] P. Hutchison、H. Maeda、C. Formby、B. J. Small、D. A. Eddins和A. C. Eddins，“声剥夺调节人类听觉脑干和皮层中的中枢增益”，听力研究，第428卷，第108683页，2023年。[6] D. Cherri、C. Formby、C. A. Secor和D. A. Eddins，“用于治疗致残性听觉过敏的过渡性干预咨询方案”，言语、语言与听力研究杂志，第67卷，第6期，第1886–1902页，2024年。[7] H. A. Hansen、A. B. Leber和Z. M. Saygin，“哪些声源会触发恐音症？不仅仅是咀嚼和呼吸”，临床心理学杂志，第77卷，第11期，第2609–2625页，2021年。[8] D. M. Orloff、D. Benesch和H. A. Hansen，“FOAMS的整理：一个免费开放获取恐音症刺激数据库”，开放心理学数据杂志，第11卷，第1期，2023年。

[9] B. Bahmei、E. Birmingham和S. Arzanpour，“使用视觉变换器进行恐音症声音识别”，载于*2023*年第*45*届*IEEE*医学与生物学工程学会国际年会（*EMBC*）。IEEE，2023年，第1–4页。[10]C. Marosv¨olgyi，“使用PaSST进行恐音症声音识别”，硕士学位论文，莱顿大学高级计算机科学研究所（LIACS），莱顿大学，2025年。 [在线]。可访问：https://theses.liacs.nl/pdf/2024-2025-MarosvlgyiCCharlotte.pdf[11] T. Wunrow，用于恐音症治疗的选择性噪声消除应用程序。密西西比州立大学，2024年。[12] A. Mesaros、T. Heittola、T. Virtanen和M. D. Plumbley，“声音事件检测：教程”，《*IEEE*信号处理杂志》，第38卷，第5期，第67–83页，2021年。[13] N. Turpault、R. Serizel、A. P. Shah和J. Salamon，“使用弱标注数据和声景合成在家庭环境中进行声音事件检测”，载于声学场景与事件检测和分类研讨会，2019年。[14] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal和M. Ritter，“Audio Set：用于音频事件的本体与人工标注数据集”，载于*2017*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2017年，第776–780页。[15] J. Salamon、D. MacConnell、M. Cartwright、P. Li和J. P. Bello，“Scaper：用于声景合成和数据增强的程序库”，载于*2017*年*IEEE*音频与声学信号处理应用研讨会（*WASPAA*）。IEEE，2017年，第344–348页。[16] S. Vitoratou、N. Uglik-Marucha、C. Hayes、M. Erfanian、O. Pearson和J. Gregory，“恐音症听觉触发因素的项目反应理论研究”，《听力学研究》，第11卷，第4期，第567–581页，2021年。[17] N. Turpault，“dcase20 task4：DCASE 2020任务4基线”，2020年，用于声景合成脚本，访问日期：2026年1月20日。[在线]。可访问：https://github.com/turpaultn/dcase20 task4[18] S. Oh、K. Palmer、D. Sabatina、A. Pietrini、C. O’Reilly和S. V. Shinkareva，“从咀嚼到啁啾：恐音症视听触发因素档案（MATA）”，*bioRxiv*，2025年。[19] E. Fonseca、X. Favory、J. Pons、F. Font和X. Serra，“FSD50K：一个人工标注声音事件的开放数据集”，《*IEEE/ACM*音频、语音与语言处理汇刊》，第30卷，第829–852页，2021年。[20] K. J. Piczak，“ESC：用于环境声音分类的数据集”，载于第*23*届*ACM*多媒体国际会议论文集，2015年，第1015–1018页。[21] Y. Gong、J. Yu和J. Glass，“V̄ocalSound：用于改进人类发声识别的数据集”，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第151–155页。[22] G. Dekkers、S. Lauwereins、B. Thoen、M. W. Adhana、H. Brouckxon、B. Van den Bergh、T. Van Waterschoot、B. Vanrumste、M. Verhelst和P. Karsmakers，“SINS数据库：利用声学传感器网络检测家庭环境中的日常活动”，载于声学场景与事件检测和分类*2017*年研讨会（*DCASE2017*）论文集，2017年，第32–36页。

[23] A. Mesaros、T. Heittola 和 T. Virtanen，《用于声学场景分类和声音事件检测的 TUT 数据库》，载于*2016*年第*24*届欧洲信号处理会议（*EU-SIPCO*）。IEEE，2016年，第1128–1132页。[24] D. Snyder、G. Chen 和 D. Povey，《MUSAN：音乐、语音和噪声语料库》，*arXiv* 预印本 *arXiv:1510.08484*，2015年。[25] 《使用 YAMNet 进行声音分类》，https://www.tensorflow.org/hub/tutorials/yamnet，访问日期：2026年1月20日。[26] T. Morocutti、F. Schmid、J. Greif、F. Foscarin 和 G. Widmer，《探索声音事件检测模型中性能与复杂度的权衡》，*arXiv* 预印本 *arXiv:2503.11373*，2025年。[27] M. Schuster 和 K. K. Paliwal，《双向循环神经网络》，*IEEE* 信号处理汇刊，第45卷，第11期，第2673–2681页，1997年。[28] K. Cho、B. Van Merri¨enboer、C. Gulcehre、D. Bahdanau、F. Bougares、H. Schwenk 和 Y. Bengio，《使用 RNN 编码器-解码器学习用于统计机器翻译的短语表示》，*arXiv* 预印本 *arXiv:1406.1078*，2014年。[29] J. Chen、X. Cai、Z. Liu、H. Zhang、L. Zuo 和 M. Wu，《基于预训练模型的半监督声音事件检测，用于 DCASE 2024 任务4》，DCASE2024 挑战赛，技术报告，2024年。[30] S. Hochreiter 和 J. Schmidhuber，《长短期记忆网络》，神经计算，第9卷，第8期，第1735–1780页，1997年。[31] H. Jaeger，《分析和训练循环神经网络的“回声状态”方法（附勘误说明）》，德国波恩：德国国家信息技术研究中心 *GMD*技术报告，第148卷，第34期，第13页，2001年。[32] H. Jaeger、M. Luko˘sevi˘cius、D. Popovici 和 U. Siewert，《带有泄漏积分神经元的回声状态网络的优化与应用》，神经网络，第20卷，第3期，第335–352页，2007年。[33] T. Akiba、S. Sano、T. Yanase、T. Ohta 和 M. Koyama，《Optuna：新一代超参数优化框架》，载于第*25*届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，2019年，第2623–2631页。[34] C.. Bilen、G. Ferroni、F. Tuveri、J. Azcarreta 和 S. Krstulovi´c，《声音事件检测稳健评估框架》，载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2020年，第61–65页。[35] F. Ronchini 和 R. Serizel，《在合成声景上评估的先进声音事件检测系统基准测试》，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第1031–1035页。[36]D. P. Kingma，《Adam：一种随机优化方法》，*arXiv* 预印本 *arXiv:1412.6980*，2014年。[37] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga等，《PyTorch：一种命令式风格、高性能的深度学习库》，神经信息处理系统进展，第32卷，2019年。

[38] V. Morfi, I. Nolasco, V. Lostanlen, S. Singh, A. Strandburg-Peshkin, L. F. Gill, H. Pamula, D. Benvent 和 D. Stowell, “少样本生物声学事件检测：DCASE 2021 挑战赛中的一项新任务。”载于第六届声学场景与事件检测和分类研讨会论文集（*DCASE 2021*），2021年，第 145–149 页。