# WESR：扩展和评估词级事件语音识别

**ChenchenYang**[1,2] **Kexin Huang**[1] **LiweiFan**[1] **Qian Tu**[1]**BotianJiang**[1,2] **Dong Zhang**[1] **Linqi Yin**[1,2] **Shimin Li**[1]**Zhaoye Fei**[1] **QinyuanCheng**[1] **XipengQiu**[1,2]*ccyang6971@gmail.com, chengqy21@m.fudan.edu.cn

[1]复旦大学[2]上海创新研究院

arXiv:2601.04508v1 [cs.CL] 8 Jan 2026

## 摘要

语音不仅传递语言信息，还包含丰富的非语言发声事件，例如笑声和哭泣。尽管语义转写已得到充分研究，但对非语言事件进行精确定位仍是一项关键且研究不足的挑战。现有方法的任务定义不够完善，类别覆盖范围有限，时间粒度也含糊不清。此外，这些方法缺少标准化的评估框架，阻碍了下游应用的发展。为弥补这一空白，我们首先构建了一个经过改进的分类体系，涵盖 21 种声音事件，并将其划分为离散型（独立出现）和连续型（与语音混合）两类。基于这一改进后的分类体系，我们推出**WESR-Bench**，这是一个由专家标注的评估集（900+ 条话语），并采用新颖的位置感知协议，将自动语音识别错误与事件检测区分开来，从而能够精确测量离散事件和连续事件的定位效果。我们还通过构建一个时长为 1,700+ 小时的语料库建立了强基线，并训练了专用模型；这些模型在保持自动语音识别质量的同时，表现优于开源音频语言模型和商业API。我们预计，WESR 将成为未来研究如何对丰富的真实世界听觉场景进行建模的重要基础资源。 [1]




图 1：我们的 WESR 任务概览，该任务生成带有明确词级对齐连续/离散事件标签的转录文本。

## 1 引言

语音是人类交流的重要组成部分，不仅传递文本信息，还包含丰富的非语言线索，例如说话者的情绪、语调，以及笑声、咳嗽和耳语等多种语音事件。这些事件反映了说话者的意图和表达风格，显著增强了语音的语境含义和表现力。

然而，仅仅知道某个事件发生了还不够；准确知道它发生在哪里同样对于理解其含义至关重要对于理解其含义同样至关重要。句子级检测常常无法捕捉词语与非言语声音之间的具体关系。例如，比较“<笑声> 我真不敢相信你竟然做了那件事！</笑声>”，这听起来像一个友善的玩笑；再比如“我真不敢相信你竟然做了 <笑声> 那件事</笑声>”，在特定词语上的笑声可能暗示说话者是在嘲讽该行为本身，或对此表示难以置信。尽管词汇内容相同，但笑声出现的位置完全改变了句子的含义。因此，我们需要进行细粒度的逐词建模，以捕捉这些细微差异，真正理解说话者的意图。

尽管如此，实现词级发声事件转录仍是一项重大挑战。传统的自动语音识别（自动语音识别）和发声事件检测（发声事件检测）系统无法可靠地完成此类任务。一方面，大多数自动语音识别模型只专注于将语音转换为纯文本，将非语言事件视为噪声而忽略，从而丢失了关键语境。另一方面，传统的发声事件检测数据集和方法采用话语级或帧级分类，缺少与口语词语的细粒度对齐。

* 通讯作者。[1] 可访问 WESR。

| 数据集 | 语言 | # 类别 | 时长（小时） | 连续性 | 多模态 | 来源 | 标注 |
|---|---|---|---|---|---|---|---|
| NonverbalTTS（Borisov 等人，2025年） | EN | 10 | 17 | ✗ | ✓ | VoxCeleb、Expresso | 流水线 |
| NVSpeech-170k（Liao 等人，2025年） | 英语，中文 | 18 | 732 | ✗ | ✓ | 米哈游，Emilia | 模型 |
| NonVerbalSpeech-38K（Ye 等人，2025年） | 英语，中文 | 10 | 131 | ✓ | ✗ | Web | 流水线 |
| SMIIP-NV (Wu 等人，2025年) | ZH | 3 | 33 | ✗ | ✓ | 内部录制 | – |
| Synparaspeech(Bai 等人，2025年) | ZH | 6 | 119 | ✗ | ✗ | 合成 | – |
| MNV-17 (Mai 等人，2025年) | ZH | 17 | 8 | ✗ | ✗ | 内部录制 | – |
| CapSpeech-SEDB（Wang 等人，2025年） | EN | 10 | <1 | ✓ | ✗ | 合成 | – |
| **WESR-Bench** | EN, ZH | **21** | **3** | ✓ | ✓ | Web | 人工 |
| **WESR-Train** | EN, ZH | **21** | **1,767** | ✓ | ✓ | Web | Gemini |

表1：现有非语言语音数据集的比较。“Lang.”表示语言覆盖范围。“# Cat.”表示包含的类别数量。“Dur.”表示总时长。“Cont.”表示数据集是否包含连续事件。“Multi.”表示话语是否可以标注多个事件类别。“–”表示标注可直接从录音或合成过程中获得。

近期研究推出了若干用于事件感知自动语音识别的词级发声事件数据集，但这些数据集要么只涵盖少量事件类别，要么音频时长有限，要么语言比例极不均衡。此外，尽管一些语料库包含连续事件标签，但它们通常将这些事件视为重叠声音事件，而非调节语音本身的发声。

在评估方面，当前的事件感知自动语音识别系统通常使用词错误率（WER）和句子级事件分类准确率进行评估，却未明确考虑事件的词级位置。少数研究进一步对句子内标签的位置错位进行惩罚，但其指标仍无法自然处理同一话语中出现多个事件的情况。此外，目前也没有针对跨越多个词的连续事件的有效评估方法。

为弥补这些不足，我们提出**词级事件语音识别（WESR）**。在评估方面，我们推出WESR-Bench，这是一个由专家标注的评估集，包含 900+ 条话语，涵盖 21 种事件类别（15 种离散类别和 6 种连续类别）；同时，我们还提出一种稳健且考虑位置的评估协议，将词汇错误与事件定位分开，并联合评估事件类型和词级对齐，原生支持在同一句话中出现多个事件。在模型构建方面，我们通过构建 WESR-Train，训练WESR模型。WESR-Train 是一个大规模语料库，包含总计 1,700+小时带有词级事件转写的语音。通过大量实验，我们证明了数据中心方法的有效性：由此训练出的模型优于强大的开源音频理解基线，也优于商业API，同时保持自动语音识别质量。总之，我们的模型和评估框架为在双语自然语音中训练和基准测试事件感知自动语音识别奠定了可靠基础。

总而言之，我们的主要贡献如下：

- **全面定义**：我们通过建立严谨的21类发声事件分类体系，并区分离散和连续属性，对词级事件-语音识别（WESR）任务进行了形式化定义，为在词汇内容之外建模副语言信息提供了全面的框架。

- **基准测试**：我们推出 WESR-BENCH，这是一个包含 900+ 个样本、由专家标注的测试集，涵盖基于我们的分类体系构建的双语自然语音。

- **面向 WESR 的强基线**：我们构建了一个专为 WESR 设计的基线模型，在 WESR 上的表现优于强大的开源音频语言模型和专有 API，可作为社区的便捷工具。

## 2 相关工作

### 2.1 自动语音识别

近年来，在大规模预训练和多任务学习的推动下，自动语音识别技术取得了显著进展。OpenAI 的 Whisper（Radford 等人，2023年）等知名模型支持多语种识别和多种任务，包括语音翻译和对齐，展现出强大的跨语言泛化能力。类似地，SeedASR（Bai 等人，2024年）以及

FireRedASR（Xu 等人，2025年）已广泛应用于工业界和学术界。尽管这些模型在语言内容识别方面性能出色，但通常会忽略语音中的非语言事件和环境信息。

### 2.2 声音事件检测

发声事件检测最初是一项分类任务，旨在确定一个音频片段是否包含特定类型的声音事件。早期数据集，如 ESC-50（Piczak，2015年）和 Vocal-Sound（Gong 等人，2022年），提供了明确的类别标签，从而支持单标签或多标签模型训练。这些任务通常侧重于话语级事件判别，难以准确识别事件的时间位置。AudioSet（Gemme-ke 等人，2017年）通过引入大规模、多标签的弱标注音频扩展了这一任务，并于 2021 年提供了带有精确时间戳的强标签，从而支持时间感知事件检测。大多数现有方法，如 HTS-AT（Chen 等人，2022a），将事件检测定义为话语级分类；而 PANNs（Kong 等人，2020年）等其他方法则支持包含时间信息的帧级输出。然而，这些方法缺乏与语义信息的整合，也未对语音内容与事件之间的关系进行联合建模。

### 2.3 非语言语音语料库

近期一系列数据集聚焦于非语言事件的内联建模。一些方法通过流水线生成词级标签（先进行帧级事件检测，再通过词对齐将事件插入转录文本），例如 NonverbalTTS（Borisov 等人，2025年）和 NonVerbalSpeech-38K（Ye 等人，2025）。这种方法能够从大规模语料库中挖掘包含事件的语音片段，但标签准确性和标签覆盖率受事件检测模型限制；多阶段处理流程（事件分类、自动语音识别和对齐）会导致误差累积，同时其性能也受标注模型本身的限制（Mai 等人，2025）。

其他方法，例如 Synparaspeech（Bai 等人，2025年）和 CapSpeech（Wang 等人，2025年），通过合成获取带有事件标注的数据。Synparaspeech 使用语音转换来转换发声事件片段，而语音部分则由文本转语音生成；CapSpeech 则依靠人工标注人员将事件音频插入语音录音中。这些方法能够实现精确的词级标签（插入位置有明确指定），但生成的音频往往不够自然。

也有一些方法依赖人工标注。NVSpeech-170k（Liao 等人，2025年）对涵盖 18 种事件类别的 48k 个音频片段进行了人工标注，随后在这一标注子集上训练自动语音识别模型，并用该模型将数据集自动扩充至 17 万个样本。SMIIP-NV（Wu 等人，2025）和 MNV-17（Mai 等人，2025）招募参与者录制带有非语言事件的语音。SMIIP-NV 收集了 33 小时、包含 3 种事件类别的人工录制数据；MNV-17 收集了 7.55 小时、包含 17 种类别的人工录制中文数据。

然而，这些方法存在若干共同局限：1）事件类别的设计要么过于精简，要么包含许多区分度不明显的标签（先前工作所用标签的汇总见补充表9）；2）用于稳健评估的高质量标注数据仍然有限；3）连续事件的定义不清晰——即使数据集包含诸如 <B>...</B>之类用于表示连续性的标签，它们通常表示重叠的声音，而非与转录文本相结合的非言语语音；4）目前尚无广受认可的评估协议：大多数事件感知自动语音识别系统仅采用分类F1和词错误率/字错误率进行评估，忽略了词级定位；而 NonVerbalSpeech-38K 中的 TPD/NTD 等指标无法处理同一句话中的多个事件或连续事件，MNV-17 中的话语级准确率则掩盖了各类别的表现，也仍无法反映词级性能。所有数据集的比较见表1。

## 3 任务定义

为促进副语言建模的标准化评估，我们定义了 WESR 任务。给定输入音频流，WESR 任务的目标是生成转录文本，其中不仅包含口语内容（与传统自动语音识别相同），还会在转录文本的正确词级位置标注非言语发声事件标签。具体而言，输出的转录文本应包括：**1) 离散事件标签：** 这些标签标记音频中特定时间点出现的简短离散事件，例如 [大笑]和 [清_嗓]，。**2)连续事件**

| WESR-Bench 中的示例 | Tags | Type |
|---|---|---|
| 好的，[清清_嗓子]，那我们先来说一个比较轻松的话题。，嗯，可能。 | [清_嗓] | 离散 |
| 亚历山大！ [大笑]哦，我的小战士。到这儿来。来吧。 [大笑] | [大笑] | 离散 |
| <笑声>诶，不是重点，然后我就想说</笑声>，那我就可以稍微减少一点点儿，就是碳水化合物，然后呢可能多吃一点儿这个蛋白质，然后多吃纤维。 | <笑声> | 连续 |
| <歌唱> 我祝你圣诞快乐，我祝你圣诞快乐，祝你新年快乐年份。 </歌唱> | <歌唱> | 连续 |
| <大喊>住手，快点住手</大喊>[咯咯笑]那我走了。 | <大喊>，[咯咯笑] | 混合 |
| <哭泣> 哦，我的脸，我的脸。 </哭泣> <大喊> 我曾经把罪带到了这个世界上。[吸气] 我不能再冒这个险了。 </大喊> [抽泣] | <哭泣>，<大喊>，[吸气]，[抽泣] | 混合 |

表2：WESR-Bench 中的示例，展示了三种事件类型：离散事件、连续事件，以及在单个话语中包含多种事件类型的混合场景。

**标签：** 这些标签描述会调节语音本身的发声事件，例如边笑边说话或歌唱。它们以如下成对标签表示：<笑着说>...</laughing>和 <歌唱>... </singing>。

在选择类别时，我们从先前研究中报告、且常见于对话类媒体（如播客、直播、有声书）中的一份宽泛发声事件列表着手。我们还通过在留出数据子集上进行试标注来进一步完善这组类别，剔除标注人员认为存在歧义或不一致的标签，最终形成一套精简而具有表达力的分类体系，既涵盖离散事件和连续事件，又适用于大规模标注和可靠评估。完整的标签分类体系详见补充表 10。

形式上，给定输入音频片段 $x$，模型会输出一个序列 $Y = (y_1, y_2, \ldots\ldots, y_n)$，其中每个 $y_i$ 要么是口语内容中的一个词，要么是代表发声事件的事件标签，并且位置精确；对于连续事件，还需进行恰当的跨度标记。例如：

输入音频：[音频片段]（一个人耳语着说“你好”，然后大笑。）输出转录文本：<耳语> 你好 </耳语> [大笑]

## 4 WESR-Bench

本节介绍 WESR-Bench，详细说明专家标注数据集的构建过程以及标准化的词级评估指标。

### 4.1 数据构建

**数据整理** 我们从多种来源收集网络规模音频数据，包括电影、电视剧、播客和有声书。为确保音频质量，我们使用 MossFormer2（Zhao 等人，2024年）进行降噪，并滤除 DNSMOS 低于 2.0 的样本，仅保留适合严格评估的高质量录音。

**混合检索策略** 由于一般语音中的发声事件十分稀少，通过随机采样构建包含丰富发声事件的评估集并不切实际。为解决这一问题，我们开发了一种混合检索机制，以高效挖掘目标样本。我们利用两种互补的模型：1）BEATs（Chen 等人，2022b），一种预训练声学编码器，可捕捉非语言音频模式，用于基于音频的向量搜索；2）AF-CLAP（Ghosh 等人，2025年），一种增强型 CLAP 风格编码器，能够实现音频事件的高效音频—文本对齐，用于对自动语音识别转录文本进行基于文本的检索。针对每个标签类别，我们分别设计了三个具有代表性的文本模态查询和三个音频模态查询。通过检索与这些查询高度相似的话语，我们获得了一个包含 1,297 条话语的候选集，以供专家标注并纳入 WESR-Bench。

### 4.2 人工标注

为确保该基准数据集能够作为可靠的黄金标准，我们采用了严格的人工标注流程。我们招募了三名标注人员，他们在参与标注前完成了强制培训计划并通过了资格测试。培训内容包括事件本体教程以及带反馈的指导练习（标注指南和界面见附录A）。标注人员按每小时标注音频 30 美元的标准获得报酬。在标注阶段，他们独立工作，不参考第4.1节中的检索结果，插入事件标签。最




图2：WESR-Bench 中的标签分布。内环展示主要事件类别，每个类别以不同颜色表示。外环展示每个类别中的具体标签实例。

后，一位资深专家审查了所有样本，以核验边界精度和分类正确性，确保质量。

### 4.3 数据集分析

**统计信息** 剔除标注者未发现有效发声事件或认为时间边界存在歧义的话语后，最终的基准数据集包含 927 条经验证的话语。标签类别分布如图 2所示，各话语的统计数据详见表 3。值得注意的是，29.13% 的话语包含多个标签，24.60% 的话语包含多种不同的事件类型，这表明单条话语中存在丰富多样的发声事件。语料库共包含 1,918 次标签出现，其中连续事件占 58.8%，离散事件占 41.2%。按时长计算，该数据集由 58% 的中文和 42% 的英语组成。

**案例研究** 表 2 展示了我们提出的 WESR-Bench 中的案例，说明了词级事件语音识别任务中事件的多样性。我们展示了离散事件，例如 [清嗓_喉咙] 和 [大笑]，，也展示了连续事件，例如 <歌唱>和<耳语>。我们还展示了一个话语中可能出现混合事件标签。这些示例展示了我们的方法能够捕捉细粒度词级边界，并为广泛的声音事件类型提供丰富的标注。

提供丰富的标注，涵盖多种声音事件类型。

### 4.4 评估协议

为了严格评估 WESR 的性能，我们聚焦于两个关键方面：是否成功检测到所有出现的发声事件，以及预测的位置是否准确。由于预测文本与真实标注不同时，直接比较事件标签可能导致对齐错误，因此我们首先在词级别对假设转录文本和参考转录文本进行对齐。随后，我们根据该对齐映射并比较事件标签，以计算最终指标，如图3所示。具体流程如下：

**步骤 1：事件保留对齐** 为对齐假设序列和参考序列，同时不更改事件标签，我们采用事件保留对齐流程。首先，暂时从参考序列中移除事件标签，以获取用于对齐的纯文本，同时保留假设序列中的事件标签。然后，使用 SequenceMatcher 将假设序列与参考序列对齐，生成编辑操作（插入、删除、替换）。最后，执行这些操作：插入操作将参考词添加到假设序列中；删除操作仅移除非事件词；对于替换操作，提取并保留被替换片段中的事件标签，将该片段替换为参考文本，并在新片段中最相似的位置重新插入提取的事件标签。该流程的伪代码见附录 B。

**步骤 2. 将事件映射到词语** 为了精确定位每个标签，我们引入“词位置”和“词间位置”的概念。连续事件会被分配给其跨度内的所有词语，表示该事件会持续跨越这些

| 指标 | # 标签 | # 话语 | % |
|---|---|---|---|
| 标签总数 | 1 | 657 | 70.87 |
| | 2 | 184 | 19.85 |
| | $\geq 3$ | 86 | 9.28 |
| 唯一标签 | 1 | 699 | 75.40 |
| | 2 | 180 | 19.42 |
| | $\geq 3$ | 48 | 5.18 |

表3：按标签统计的话语分布。**标签总数**：话语中的所有事件标签（连续对< >...</> 计为一个）。**唯一标签**：不同的事件类别（重复标签只计一次）。“# Utt.”表示话语数量。

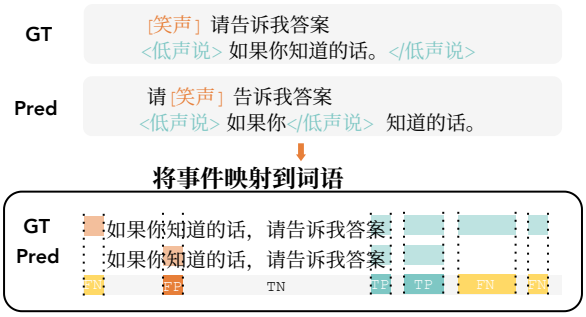


图3：事件保留对齐后，WESR评估方法步骤2和步骤3的示意图。提取事件标签，并将其映射到相应的词位置或词间位置，以便计算指标。

词；离散事件被分配到两个词之间的位置，用于标记瞬时事件的发生。因此，对于由 $N$ 个词组成的序列，共有 $2N+1$ 个可能的位置： $N$ 个词位置和 $N+1$ 个词间位置（包括起始和结束位置）。这种设计既能捕捉连续区间，也能精确定位插入点，从而统一评估不同类型的标签。

**步骤3：指标计算** 根据对齐后的序列以及统一的词位置/词间位置，我们计算每种标签类型的真正例（$TP$）、假正例（$FP$）和假负例（$FN$）。 $TP$表示正确标注了该事件的位置；$FP$表示错误标注了不应出现事件的位置；$FN$ 表示应标注事件标签但缺失或标注不正确的位置。请注意，我们不统计真负例，因为对话语音中的大多数位置都不包含发声事件。

## 5 为 WESR 构建强基线

本节将通过构建 WESR-TRAIN数据集并训练WESR专用模型，为 WESR 提出一个强基线。随后，我们将在 WESR-Bench 上，将这些模型与近期的开源音频语言模型及商业 API 进行评估比较。

### 5.1 WESR-Train

为促进词级事件语音识别的训练，我们构建了一个名为 WESR-TRAIN 的大规模弱标注数据集。

**数据收集与标注** 对于从网页获取的数据，我们采用与第 4.1 节所述相似的数据整理和混合检索策略。我们使用与构建 WESR-Bench 时不同的文本/音频查询来检索多样化的样本。为确保评估公平，我们将收集到的数据与 WESR-Bench 进行去重，以避免任何重叠。随后，使用 Gemini 自动标注收集到的数据 [2]，（数据收集时可用的最新版本），以生成词级事件标注（提示词详情见附录D ）。

**从其他数据集适配** 对于开放源代码数据集，我们纳入了来自 NonverbalTTS、NVSpeech-170k、NonVerbalSpeech-38K 和 SMIIP-NV 的词级人声事件数据，并将其标注转换为我们的格式。我们通过仔细映射，将它们的人声事件标签标准化，使其符合 WESR 分类体系，并移除带有分类体系之外标签的音频。为确保映射质量，我们对映射后的标注进行人工审查，尤其关注 NonVerbalSpeech-38K，因为它采用的连续事件定义与我们的不同。我们还发现，标注为连续<B>...</B> 标签的 NonVerbalSpeech-38K 音频存在大量标注错误，因此将其排除在训练集之外。WESR-Train 的最终数据分布如表5所示。

### 5.2 跨骨干模型适配

借助 WESR-Train 提供的丰富标注，我们对三个不同的骨干模型进行监督微调：Whisper-Large-v3 (1.5B) (Radford 等人，2023年)、Kimi-Audio-7B- Instruct (7B) (KimiTeam 等人，2025年) 和 Qwen3- Omni-Instruct (30B) (徐 等人，2025a)，并在 WESR-Bench 上评估其性能。我们的实验设置详见附录C。

如表4所示，尽管模型规模（1.5B 至 30B 参数）差异显著，我们的方法在不同架构上均取得了稳定的性能，宏平均F1分数介于 37.7% 至 38.0% 之间。这表明，我们的训练方法能够很好地泛化至不同的模型架构和规模。

在后续评估与比较中，我们选择微调后的 Qwen3-Omni 作为代表模型，因为它在不同事件类别上始终表现出色。

### 5.3 与其他基线的比较

为全面评估我们提出的方法，我们将其性能与

[2] gemini-2.5-pro

| Tag | Kimi-Audio P / R / F1 (%) | Qwen3-Omni P / R / F1 (%) | Gemini-2.5-Pro P / R / F1 (%) | Gemini-3-Pro P / R / F1 (%) | WESR-Whisper P / R / F1 (%) | WESR-Kimi P / R / F1 (%) | WESR-Qwen P / R / F1 (%) |
|---|---|---|---|---|---|---|---|
| <哭泣> | 28.9 / 04.6 / 08.0 | 55.3 / 11.7 / 19.4 | 59.4 / 80.3 / 68.3 | 66.2 / 73.7 / 69.7 | 65.1 / 89.4 /**75.3** | 60.7 / 90.2 / 72.5 | 60.8 / 92.1 / 73.3 |
| <笑声> | 00.4 / 00.3 / 00.4 | 03.8 / 12.5 / 05.8 | 22.5 / 17.3 / 19.6 | 30.8 / 25.6 /**28.0** | 29.8 / 12.5 / 17.6 | 26.3 / 15.7 / 19.7 | 35.4 / 22.4 / 27.5 |
| <喘气> | 0 / 0 / 0 | 0 / 0 / 0 | 39.8 / 30.5 / 34.5 | 37.8 / 20.5 / 26.6 | 32.4 / 32.9 / 32.7 | 34.5 / 24.5 / 28.6 | 38.0 / 36.9 / **37.5** |
| <大喊> | 31.5 / 17.2 / 22.2 | 73.0 / 41.6 / 53.0 | 77.9 / 53.7 / 63.6 | 62.8 / 68.4 / 65.5 | 72.1 / 61.7 /**66.5** | 65.4 / 55.7 / 60.2 | 72.5 / 59.2 / 65.2 |
| <歌唱> | 57.1 / 00.5 / 00.9 | 96.3 / 75.9 / 84.9 | 98.5 / 87.2 / 92.5 | 97.3 / 81.6 / 88.7 | 99.6 / 92.4 / 95.9 | 99.4 / 93.7 /**96.5** | 98.8 / 93.0 / 95.8 |
| <耳语> | 0 / 0 / 0 | 59.5 / 10.8 / 18.3 | 96.4 / 14.1 / 24.7 | 94.0 / 37.1 / 53.2 | 85.6 / 64.4 / 73.5 | 84.2 / 67.7 / 75.1 | 85.9 / 69.9 /**77.1** |
| [呼吸] | 10.3 / 06.2 / 07.8 | 03.3 / 02.1 / 02.6 | 05.0 / 02.1 / 02.9 | 16.4 / 22.9 / **19.1** | 0 / 0 / 0 | 03.6 / 02.1 / 02.6 | 02.3 / 02.1 / 02.2 |
| [轻笑] | 0 / 0 / 0 | 40.0 / 03.1 / 05.7 | 12.8 / 32.3 / 18.3 | 20.0 / 12.3 / 15.2 | 16.8 / 52.3 / 25.5 | 20.9 / 52.3 /**29.8** | 16.6 / 53.8 / 25.4 |
| [清嗓子_] | 38.9 / 48.3 / 43.1 | 30.4 / 48.3 / 37.3 | 44.2 / 65.5 / 52.8 | 51.4 / 62.1 / 56.2 | 67.9 / 65.5 / 66.7 | 74.1 / 69.0 / **71.4** | 65.6 / 72.4 / 68.9 |
| [咳嗽] | 39.7 / 31.3 / 35.0 | 64.6 / 42.4 / 51.2 | 51.4 / 55.6 / 53.4 | 65.4 / 52.0 / 58.0 | 74.4 / 67.7 /**70.9** | 71.8 / 56.6 / 63.3 | 69.6 / 55.6 / 61.8 |
| [人群_笑声] | 0 / 0 / 0 | 0 / 0 / 0 | 66.7 / 15.4 / 25.0 | 29.0 / 23.1 / 25.7 | 42.9 / 15.4 / 22.6 | 44.4 / 10.3 / 16.7 | 56.2 / 23.1 / **32.7** |
| [哭泣] | 11.4 / 08.3 / 09.6 | 00.5 / 02.1 / 00.9 | 16.2 / 22.9 /**19.0** | 05.9 / 02.1 / 03.1 | 100 / 08.3 / 15.4 | 83.3 / 10.2 / 18.2 | 28.6 / 04.2 / 07.3 |
| [呼气] | 03.2 / 03.1 / 03.2 | 10.0 / 28.1 / 14.8 | 06.5 / 06.2 / 06.3 | 14.3 / 09.4 / 11.3 | 20.0 / 09.4 / 12.8 | 31.6 / 18.8 / **23.5** | 15.0 / 09.4 / 11.5 |
| [咯咯笑] | 16.7 / 10.7 / 13.0 | 02.4 / 07.1 / 03.5 | 10.5 / 28.6 / 15.4 | 20.0 / 22.2 / 21.1 | 25.6 / 39.3 / 31.0 | 28.6 / 35.7 /**31.7** | 19.6 / 32.1 / 24.3 |
| [吸气] | 10.2 / 07.3 / 08.5 | 11.1 / 01.5 / 02.6 | 08.7 / 09.5 / 09.1 | 06.0 / 33.6 / 10.1 | 09.6 / 06.6 / 07.8 | 13.5 / 09.4 / 11.1 | 14.8 / 09.5 /**11.6** |
| [大笑] | 20.2 / 58.8 / 30.0 | 33.3 / 38.2 / 35.6 | 33.8 / 35.9 / 34.8 | 30.8 / 61.8 / 41.1 | 45.5 / 42.0 / 43.7 | 49.6 / 45.6 /**47.5** | 44.5 / 40.5 / 42.4 |
| [咆哮] | 0 / 0 / 0 | 0 / 0 / 0 | 16.7 / 25.0 / 20.0 | 0 / 0 / 0 | 50.0 / 25.0 /**33.3** | 50.0 / 25.0 / **33.3** | 20.0 / 25.0 / 22.2 |
| [尖叫] | 04.2 / 10.0 / 05.9 | 0 / 0 / 0 | 10.0 / 20.0 / 13.3 | 08.7 / 20.0 / 12.1 | 21.1 / 40.0 / 27.6 | 08.3 / 10.0 / 09.1 | 23.5 / 40.0 /**29.6** |
| [呼喊] | 10.0 / 27.3 / 14.6 | 02.4 / 09.1 / 03.7 | 20.0 / 13.6 / 16.2 | 25.0 / 22.7 /**23.8** | 40.0 / 09.1 / 14.8 | 40.0 / 09.1 / 14.8 | 50.0 / 13.6 / 21.4 |
| [叹气] | 07.4 / 05.9 / 06.6 | 09.9 / 23.5 / 13.9 | 12.1 / 47.1 / 19.3 | 26.8 / 44.1 / 33.3 | 27.0 / 58.8 / 37.8 | 35.9 / 67.6 /**46.9** | 27.3 / 61.8 / 37.8 |
| [啜泣] | 05.4 / 19.3 / 08.5 | 00.6 / 03.5 / 01.0 | 12.8 / 40.4 / 19.5 | 09.6 / 42.1 / 15.7 | 12.6 / 50.9 / 20.2 | 14.1 / 48.2 / 21.9 | 13.7 / 57.9 /**22.1** |
| 微平均 | 19.4 / 04.5 / 07.3 | 47.8 / 27.0 / 34.5 | 64.3 / 46.4 / 53.9 | 63.8 / 54.7 / 58.9 | 72.5 / 68.9 / 70.6 | 71.3 / 69.7 / 70.5 | 71.2 / 71.7 /**71.4** |
| 宏平均 | 14.1 / 12.3 / 10.3 | 23.6 / 17.2 / 16.9 | 34.4 / 33.5 / 29.9 | 34.2 / 35.1 / 32.3 | 44.7 / 40.2 / 37.7 | 44.8 / 38.9 / 37.8 | 40.9 / 41.6 /**38.0** |

表4：各模型在 WESR-Bench 的所有发声事件类别上的性能。前四列展示了使用两样本提示的结果（Kimi-Audio、Qwen3-Omni、Gemini-2.5-Pro、Gemini-3-Pro），后三列展示了我们的 WESR训练模型（WESR-Whisper、WESR-Kimi、WESR-Qwen）的结果。“微平均”通过对所有样本直接求平均计算得出，“宏平均”则通过对各类别的性能求平均计算得出。每行中最佳和次佳的 F1分数分别以粗体和下划线标出。

| 数据来源 | 时长（小时） | 语言 |
|---|---|---|
| NonverbalTTS | 14 | EN |
| NVSpeech-170k | 332 | 英语、中文 |
| NonVerbalSpeech-38K | 87 | 英语、中文 |
| SMIIP-NV | 35 | ZH |
| Gemini-标注 | 1,299 | 英语、中文 |
| WESR-TRAIN | **1,767** | EN, ZH |

表 5：WESR-Train数据集的组成。“Dur.”表示总时长。

若干具有代表性的音频语言模型（音频语言模型）在 WESR-Bench 上的其他基线。我们在评估中考察的音频语言模型如下：1）最新的开源音频语言模型，例如 Kimi-Audio（KimiTeam 等人，2025年）、MiMo-Audio（张等人，2025年）和 Qwen3-Omni（徐等人，2025a）。2）商业 API，例如 Gemini [3] 和 GPT [4]。我们使用相同的指令提示这些模型（见附录 D），其中包含两个上下文示例，说明离散标签和连续标签的用法。详细结果见表4。

[3]gemini-2.5-pro, gemini-3-pro[4] gpt-4o-audio-preview

如表 4所示，不同模型在非语言事件检测方面表现出显著差异。总体而言，我们的微调版 Qwen3-Omni WESR 模型在所有类别中均取得最高的 F1分数，宏平均F1 达到 38.0%，较 Qwen3-Omni 绝对提升 21.1%，较 Gemini-2.5-Pro 提升 8.1%。这一提升在 <喘气>、 <耳语>以及各种笑声相关标签（[轻笑]、[咯咯笑]、[大笑]）等具有挑战性的类别中尤为明显：少样本模型往往几乎无法召回这些事件，而微调模型的检测率则显著更高。例如，WESR-Whisper 对 <耳语> 的召回率为 64.4 %，而 Gemini-3-Pro 为 37.1%；对 [轻笑] 的召回率为 52.3%，而 Gemini-3-Pro 为 12.3%。这些结果表明，尽管大型多模态模型通过提示具备一定的语音事件识别能力，但要在各种副语言现象中实现稳健、可靠的性能，系统地接触标注训练数据仍至关重要。

在评估过程中，我们还注意到，GPT-4o 和 MiMo-Audio 经常拒绝响应 WESR提示词。即使在少数生成了输出的情况下，它也未能在转录文本中插入任何事件标签

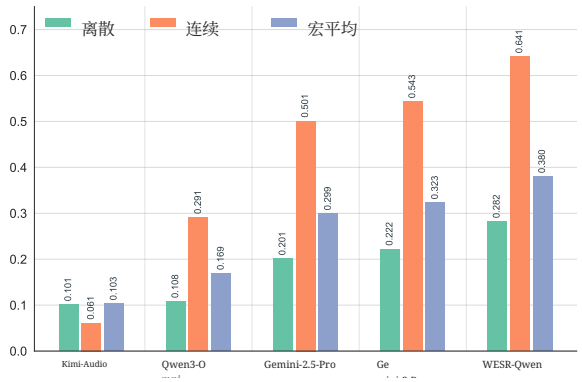


图4：WESR-Bench 上连续事件标签与离散事件标签的性能比较。

，因此不适用于我们的 WESR任务。

我们还在补充表格 8 中报告了聚合结果，其中细粒度标签被归入更宽泛的发声事件类别（例如，将 [清嗓子_和] 咳嗽归为 [咳嗽]），详细内容见补充表格10。聚合指标进一步证实了我们的模型区分不同发声事件类别的能力。我们还观察到各模型的性能趋势高度一致：所有模型在 歌唱（F1> 94%）和 耳语（F1 >75%）等明显且持续的事件上表现最佳，而在 呼吸 等细微、低能量的声学特征上则始终面临挑战。

### 5.4 离散事件与连续事件的比较

连续事件（例如 <哭泣>. . .</哭泣>）比离散事件（例如 [咳嗽]）更具挑战性。模型不仅必须推断出正确的事件类型，还必须确定准确的时间跨度边界。为研究所有基线模型对离散事件和连续事件的识别能力，我们分别计算这两类事件的 F1分数。请注意，连续事件的 F1分数通常高于离散事件：当边界略有偏差时，连续事件标签可通过词元级重叠获得部分得分（见第 4.4 节），而离散事件标签则要求位置完全匹配。

如图 4所示，开源模型表现明显不佳：Kimi-Audio 的 F1 值仅为 0.101 / 0.061（离散 / 连续），而 Qwen3-Omni 为 0.108 / 0.291。专有模型表现更好，Gemini-2.5-Pro 为 0.201 / 0.501，Gemini-3.0-Pro 为 0.222 / 0.543。WESR-Qwen 在两种事件类型上均显著优于所有基线，在离散事件上的 F1 值达到 0.282（比 Gemini-3.0-Pro 高 +27 %），在连续事件上的 F1 值达到 0.641（提升 +18 %）。这表明，我们的专门训练能有效提升精确的事件定位和准确的片段边界预测能力。

### 5.5 对自动语音识别性能的影响

为了探究 WESR 对底层音频语言模型自动语音识别能力的影响，我们在 Common Voice 15（Ardila 等人，2020年）测试集上开展实验。该测试集是广泛使用的大规模自动语音识别基准数据集。我们比较了原始模型和经过 WESR 微调的模型在英语（en）和中文（zh-CN）测试集上的词错误率（WER）。由于 Common Voice 不包含发声事件标注，因此我们在计算词错误率时排除人声事件标签，以确保比较公平。结果见表6。

| 模型 | en | 中文-CN |
|---|---|---|
| Qwen3-Omni | 7.2 | 6.0 |
| WESR-Qwen | 8.6 | 7.2 |

表 6：原始 Qwen3-Omni 与我们经过 WESR 微调的模型在 CommonVoice 15 测试集（英语和中文-CN）上的词错误率（WER，%）比较。词错误率越低，性能越好。

结果表明，我们经过 WESR 微调的模型仅使词错误率略有上升，同时保持了有竞争力的自动语音识别性能，证明 WESR 可以集成到模型中，而不会显著影响转录准确率。

## 6 结论

在这项工作中，我们提出了词级事件语音识别（WESR）任务。我们建立了一套严谨的分类体系，将 21 种发声事件划分为离散和连续类型，为该任务构建了全面的框架。我们开发了 WESR-BENCH，这是一个由专家标注的基准，采用新颖的位置感知协议，将自动语音识别错误与事件检测区分开来，从而实现精确的定位测量。我们通过构建 WESR-Train 并训练专用模型，为 WESR 建立了强基线。这些模型在不同参数规模下，均优于开源音频语言模型和商业 API，同时保持了自动语音识别质量。我们相信，WESR 将为未来研究丰富的真实世界听觉场景建模提供基础资源。

## 局限性

我们承认，本工作可能存在以下局限：1）语言覆盖范围：目前，我们的系统仅在英语和中文上经过验证。尽管底层方法与语言无关，但其在形态结构不同或资源较少的语言上的泛化能力仍未得到验证。2）资源密集度：我们高质量数据集的构建依赖商业 Gemini API 和专家标注。虽然这确保了数据质量，但也需要在成本效益和可扩展性之间进行权衡，这可能会给资源受限环境中的可复现性带来挑战。

## 伦理考量

在数据构建和建模流程中，我们始终将伦理标准放在首位。我们的数据集依据合理使用原则使用公开可用音频。我们通过提供公平报酬（30美元/小时）保障标注人员的权益。由于标注规模庞大，无法对所有数据实例进行全面的人工核验。因此，数据集可能会无意中包含不当内容。我们强调，源音频或标注中出现的任何内容均不代表作者的观点、信念或认可。我们发布代码、数据和模型仅供学术用途。

## 参考文献


Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis Tyers 和 Gregor Weber。2020年。Common Voice：一个大规模多语言语音语料库。载于第十二届语言资源与评估会议论文集，第4218–4222页，法国马赛。欧洲语言资源协会。

Bingsong Bai、Qihang Lu、Wenbing Yang、Zihan Sun、Yueran Hou、Peilei Jia、Songbai Pu、Ruibo F u、Ying- ming Gao、Ya Li 等 1 位作者。2025年。Synparaspeech：用于语音生成与理解的副语言数据集自动合成。*arXiv* 预印本 *arXiv:2509.14946*。

Ye Bai、Jingping Chen、Jitong Chen、Wei Chen、Zhuo Chen、Chuang Ding、Linhao Dong、Qianqian Dong、Yujiao Du、Kepan Gao及其他1位。2024年。Seed-自动语音识别：利用基于大语言模型的语音识别理解多样化的语音及语境。*arXiv*预印本 *arXiv:2407.04675*。

Maksim Borisov、Egor Spirin和Daria Diatlova。2025年。NonverbalTTS：一个公开的英语语料库，包含带有情感标注的文本对齐的非语言发声，用于文本转语音。*arXiv*预印本 *arXiv:2507.13155*。

Ke Chen、Xingjian Du、Bilei Zhu、Zejun Ma、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov。2022a。Hts- at：用于声音分类与检测的分层令牌语义音频变换器。载于*ICASSP 2022-2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第 646–650 页。IEEE。

Sanyuan Chen、Yu Wu、Chengyi Wang、Shujie Liu、Daniel Tompkins、Zhuo Chen 和 FuruWei。2022b。BEATs：使用声学分词器进行音频预训练。*arXiv* 预印本 *arXiv:2212.09058*。

Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。2017年。Audio set：一种用于音频事件的本体和人工标注数据集。载于*2017*年 *IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 776–780 页。IEEE。

Sreyan Ghosh、Zhifeng Kong、Sonal Kumar、S Sakshi、Jaehyeon Kim、Wei Ping、Rafael Valle、Dinesh Manocha 和 Bryan Catanzaro。2025年。Audio Flamingo 2：一种具备长音频理解和专家级推理能力的音频语言模型。*arXiv* 预印本 *arXiv:2503.03983*。

Yuan Gong、Jin Yu 和 James Glass。2022年。Vocal-sound：一个用于提升人类发声识别能力的数据集。载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 151–155 页。IEEE。

KimiTeam、Ding Ding、Zeqian Ju、Yichong Leng、Songxiang Liu、Tong Liu、Zeyu Shang、Kai Shen、Wei Song、Xu Tan、Heyi Tang、Zhengtao Wang、Chu Wei、Yifei Xin、Xinran Xu、JianweiYu、Yutao Zhang、Xinyu Zhou、Y. Charles 及另外 21 位作者。2025年。Kimi 音频技术报告。预印本，arXiv:2504.18425。

Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley。2020年。PANNs：用于音频模式识别的大规模预训练音频神经网络。*IEEE/ACM* 音频、语音与语言处理汇刊，28：2880–2894。

Huan Liao、Qinke Ni、Yuancheng Wang、Yiheng Lu、Haoyue Zhan、Pengyuan Xie、Qiang Zhang 和 Zhizheng Wu。2025年。Nvspeech：用于类人语音建模及副语言发声的集成式可扩展流程。*arXiv* 预印本 *arXiv:2508.04195*。

Jialong Mai、Jinxin Ji、Xiaofen Xing、Chen Yang、Weidong Chen、Jingyuan Xing 和 Xiangmin Xu。2025年。MNV-17：一个用于语音中非语言发声识别的高质量表演型普通话数据集。*arXiv* 预印本 *arXiv:2509.18196*。

Karol J Piczak。2015年。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，第 1015–1018 页。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brock- man、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习会议，第 28492–28518 页。PMLR。

Helin Wang、Jiarui Hai、Dading Chong、Karan Thakkar、Tiantian Feng、Dongchao Yang、Junhyeok Lee、Thomas Thebaud、Laureano Moro Velazquez、Jesus Villalba 及另外 1 位作者。2025年。CapSpeech：推动风格描述文本转语音中的下游应用。*arXiv* 预印本 *arXiv:2506.02863*。

Zhuojun Wu、Dong Liu、Juan Liu、Yechen Wang、Linxi Li、Liwei Jin、Hui Bu、Pengyuan Zhang 和 Ming Li。2025年。SMIIP-NV：面向基于大语言模型的语音合成的普通话多标注非言语表现性语音语料库。载于第 *33* 届 *ACM* 国际多媒体会议论文集，第 12564–12570 页。

Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu、Yuanjun Lv、Yongqi Wang、Dake Guo、He Wang、Linhan Ma、Pei Zhang、Xinyu Zhang、Hongkun Hao、Zishan Guo 及另外19位作者。2025年a。Qwen3-Omni 技术报告。*arXiv* 预印本 *arXiv:2509.17765*。

Kai-Tuo Xu、Feng-Long Xie、Xu Tang 和 Yao Hu。2025年b。FireRedASR：从编码器-解码器到大语言模型集成的开源工业级普通话语音识别模型。*arXiv* 预印本 *arXiv:2501.14350*。

Runchuan Ye、Yixuan Zhou、Renjie Yu、Zijian Lin、Ke-han Li、Xiang Li、Xin Liu、Guoyang Zeng 和 Zhiy-ong Wu。2025年。一种用于实现非语言语音生成与理解的可扩展流程。*arXiv* 预印本 *arXiv:2508.05385*。

Dong Zhang、Gang Wang、Jinlong Xue、Kai Fang、Liang Zhao、Rui Ma、Shuhuai Ren、Shuo Liu、Tao Guo、Weiji Zhuang 等共 12 位作者。2025年。Mimo-audio：音频语言模型是少样本学习器。*arXiv* 预印本 *arXiv:2512.23808*。

Shengkui Zhao、Yukun Ma、Chongjia Ni、Chong Zhang、Hao Wang、Trung Hieu Nguyen、Kun Zhou、Jia Qi Yip、Dianwen Ng 和 Bin Ma。2024年。Mossformer2：结合 Transformer 与无 RNN 的循环网络，实现增强型时域单声道语音分离。载于 *ICASSP 2024-2024 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 10356–10360 页。IEEE。

## A 注释详情

表 7展示了声音事件分类体系的说明以及提供给标注人员的示例。图 6 展示了注释页面。

## B 评估详情

**算法1**事件保留对齐

**要求：** 假设序列 $H$，参考序列—列 $R$

**确保：** 对齐后的假设序列 $H'$

$R_{text} \leftarrow$ 移除其中的所有事件标签 $R$

$H_{split} \leftarrow$ 将 $H$ 拆分为列表（保留事件标签作为独立元素）

$R_{split} \leftarrow$ 将 $R_{text}$ 拆分为列表

$ops \leftarrow$SequenceMatcher($H_{split}$, $R_{split}$) {获取编辑操作}

$H' \leftarrow H_{split}$

**对于**每个操作 $op$在 $ops$ **执行**

  **如果** $op$为插入**则**

    从 $R_{split}$中插入相应的词语到 $H'$ 中

  **否则，如果** $op$ 为删除 **则**

    仅从 $H'$ 中删除非事件词

  **否则，如果** $op$是替换，**则**

    $events \leftarrow$ 从替-换后的片段中提取事件标签 $H'$

    将 $H'$ 中的片段替换为对应的文本，来自 $R_{split}$

    在最相似的位置重新插入 $events$ 在新片段中

  **结束条件判断**

**结束循环**

**返回** $H'$ {对齐序列匹配参照——非事件内容}

## C 训练详情

对于 Whisper，我们使用学习率 $1 \times 10^{-5}$ 对 Whisper-Large-v3 进行微调，全局批次大小为 8，预热步数比例为 0.1，并在 H100*8 上训练 3 个训练轮次，共耗时 4 小时。

对于 Kimi，我们使用学习率 $1 \times 10^{-6}$ 对 Kimi-Audio-7B-Instruct 进行微调，打包序列长度为 4,096 个词元，并在 H100*8 上训练 3 个训练轮次，共耗时 5 小时。

对于 Qwen3-Omni，我们使用学习率 $1 \times 10^{-6}$、4,096 个词元的打包序列长度，在 H200*8 上训练 5 小时，共训练 3 个训练轮次，对 Qwen3-Omni-30B-A3B-Instruct 进行了微调。

所有超参数均通过在验证集上使用三种不同学习率进行初步实验确定，该验证集从训练集中随机采样得到。声音事件标签被添加为特殊词元。

## D 提示词

""" 您是一名音频特殊模式识别专家。请执行以下任务：转录给定音频中的文本，并根据音频中检测到的内容，将以下细粒度人声事件标签插入文本中的精确位置。支持的标签包括： [笑声]，[轻笑]，[咯咯笑]、 <笑声></笑声>、 [人群_笑声]，[哭泣]，[抽泣]、 <哭泣></哭泣>、 [咳嗽]，[清嗓_喉咙]，[呼喊]，[尖叫]，[咆哮]、 <大喊></ 大喊>、 <耳语></耳语>、 [吸气]，[呼气]、 <喘气></喘气>、 [呼吸]，[叹气]、 <歌唱></歌唱>。标签使用指南：方括号 [ ]：在事件发生的精确位置插入标签，通常位于词语之间或句子内部。例如：[吸气] 我不认为他们看到了我们。尖括号 < >...</>：使用这些标签包围在发声事件发生时说出的特定词语或短语。例如：<笑声> 就是这样！ </笑声>标注指南：仅在发声事件发生的精确位置添加事件标签。标注应精确对应事件发生的位置。如果没有需要标注的事件，输出应仅包含自动语音识别文本。音频中可能不存在任何发声事件。音频中没有发声事件时，不要插入事件标签。最终输出只能包含带有事件标注的文本。不要包含任何其他说明或格式。示例：“我不认为他们看到了我们。 [吸气] 我们继续前进吧。”或“<笑声> 就是这样！ </笑声>” """

## E 补充表格

| Tag | 说明 | 标注示例 | 音频示例 |
|---|---|---|---|
| [大笑] | 笑声 | 然后 [大笑] 那真是太棒了。 | (省略) |
| [轻笑] | 轻声笑 | 我就知道，对吧？ [轻笑] 我想的正是这样 我本来就这么想。 | （省略） |
| [咯咯笑] | 因觉得有趣而轻声、高声地笑起来 | 她把这个故事讲给我听，我就 [咯咯笑] 忍不住一直笑。 | （省略） |
| <笑声> | 一边说话一边笑 | 后来 we 甚至没有 need a 模型；<笑着>就是这样。 | （省略） |
| [人群_笑声] | 多人发出的笑声 | 他总是打我..... [人群_笑声] 哦 我的..... [人群_笑声] ··· 家长-教师 会议 were the 最难的... [人群_笑声]。 | （省略） |
| [哭泣] | 可听见的哭声 | 看看这条裙子..... [哭泣] 它看起来糟透了；我不会穿它。 | （省略） |
| [啜泣] | 断断续续的啜泣声 | 我简直不敢相信这事竟然发生了..... [啜泣] 我只是不能。 | （省略） |
| <哭泣> | 哭泣着说话／带着哭腔声音 | <哭泣>我在利丁街附近捡到了它；那是在他们婚礼当天。 | （省略） |
| [咳嗽] | 咳嗽声 | 谢谢 [咳嗽] 这两天假期。 | （省略） |
| [清_嗓] | 清嗓声 | [清_嗓子] 这场演讲令人鼓舞；让我们都来分享我们的想法。 | （省略） |
| [尖叫] | 高亢的大声发声 | 救救我！ [尖叫]请哪位来帮帮我！ | （省略） |
| [咆哮] | 低沉轰鸣的发声 | 比赛开始！ [咆哮]我们上吧！ | （省略） |
| [呼喊] | 大声呼喊或叫喊 | [呼喊]是谁？开门！ | （省略） |
| <大喊> | 一边大喊一边说话 | <大喊>相信我！求求你！你必须相信我！ | （已省略） |
| [呼吸] | 呼吸或喘息声 | [呼吸] 我欠你太多了.....只有你一无所有..... | （省略） |
| [吸气] | 可听见的吸气 | 我欠你太多了 [吸气] ..... 只有你 [吸气] 一无所有..... | （省略） |
| [呼气] | 可听见的呼气 | 我欠你太多了..... [呼气] 只有你..... [呼气] 一无所有..... | （省略） |
| <喘气> | 说话时急促呼吸 | <喘气>我已经跑不动了..... 我真的撑不下去了。 | （省略） |
| [叹气] | 叹气声 | 他痛恨这种遗传病..... [叹气] 这实在太不公平了。 | （省略） |
| <耳语> | 低声说话 | <耳语>我的英语太差了——连简单的短语都说不流利。 | （省略） |
| <歌唱> | 歌唱声 | <歌唱>你把我写进剧本，还说你一定是主角。 | （省略） |

表7：非言语发声事件的标注指南。区间事件使用尖括号 <...>，尖峰事件使用方括号 [...]。

| 聚合标签 | Kimi-Audio P / R / F1（%） | Qwen3-Omni P / R / F1（%） | Gemini-2.5-Pro P / R / F1（%） | Gemini-3-Pro P / R / F1（%） | WESR-Whisper P / R / F1（%） | WESR-Kimi P / R / F1（%） | WESR-Qwen P / R / F1（%） |
|---|---|---|---|---|---|---|---|
| 呼吸 | 22.6 / 07.8 / 11.6 | 25.7 / 13.7 / 17.9 | 30.8 / 30.9 / 30.9 | 16.6 / 35.2 / 22.5 | 33.6 / 30.7 / 32.1 | 37.9 / 27.8 / 32.0 | 37.0 / 33.8 / 35.3 |
| 咳嗽 | 51.8 / 44.5 / 47.9 | 60.7 / 50.8 / 55.3 | 56.2 / 64.1 / 59.9 | 70.5 / 62.2 / 66.1 | 77.2 / 68.8 / 72.7 | 79.4 / 63.3 / 70.4 | 71.8 / 61.7 / 66.4 |
| 哭泣 | 18.9 / 04.6 / 07.4 | 68.9 / 33.9 / 45.5 | 56.4 / 78.8 / 65.8 | 62.0 / 74.4 / 67.7 | 61.7 / 87.8 / 72.5 | 59.2 / 90.1 / 71.5 | 57.7 / 90.4 / 70.4 |
| 笑声 | 24.1 / 27.8 / 25.8 | 11.1 / 24.1 / 15.2 | 39.1 / 40.3 / 39.7 | 42.5 / 45.3 / 43.8 | 50.4 / 43.3 / 46.6 | 48.9 / 42.8 / 45.6 | 48.7 / 47.8 / 48.2 |
| 呼喊 | 30.2 / 18.0 / 22.6 | 64.3 / 41.2 / 50.2 | 77.1 / 52.0 / 62.1 | 61.0 / 66.3 / 63.5 | 72.3 / 61.0 / 66.2 | 65.0 / 54.4 / 59.2 | 72.2 / 58.8 / 64.8 |
| 歌唱 | 57.1 / 00.5 / 00.9 | 96.4 / 75.6 / 84.8 | 98.6 / 87.2 / 92.6 | 97.3 / 81.8 / 88.8 | 99.8 / 92.6 / 96.1 | 99.4 / 93.7 / 96.4 | 98.8 / 92.8 / 95.8 |
| 耳语 | 0 / 0 / 0 | 59.5 / 10.8 / 18.3 | 96.6 / 14.2 / 24.7 | 94.1 / 37.1 / 53.2 | 85.6 / 64.5 / 73.6 | 84.2 / 67.7 / 75.1 | 85.9 / 70.0 / 77.2 |
| 微平均 | 26.2 / 05.8 / 09.4 | 58.0 / 32.7 / 41.8 | 68.2 / 48.3 / 56.6 | 66.5 / 57.0 / 61.4 | 75.0 / 71.0 / 73.0 | 73.6 / 71.9 / 72.8 | 73.5 / 73.7 / 73.6 |
| 宏平均 | 29.3 / 14.7 / 16.6 | 55.2 / 35.7 / 41.0 | 65.0 / 52.5 / 53.7 | 63.4 / 57.5 / 58.0 | 68.6 / 64.1 / 65.7 | 67.7 / 62.8 / 64.3 | 67.5 / 65.0 / 65.4 |

表8：各模型在 WESR-Bench 汇总的声音事件类别上的性能。前四列显示少样本提示的结果（Kimi-Audio、Qwen3-Omni、Gemini-2.5-Pro、Gemini-3-Pro），后三列显示在我们的WESR-Train语料库上微调的模型（WESR-Whisper、WESR-Kimi、WESR-Qwen）的结果。

| 数据集 | Tags | 数量 |
|---|---|---|
| NonverbalTTS | [喘气], [咕哝], [吸鼻子], [清嗓子], [呻吟], [叹气], [打鼾], [咳嗽], [笑声], [打喷嚏] | 10 |
| NVSpeech-170k | [呼吸], [笑声], [确认-en], [嗯], [叹气], [惊讶-ah], [惊讶-oh], [不满-hnn], [惊讶-wa], [疑问-yi], [疑问-ei], [咳嗽], [疑问-ah], [疑问-oh], [惊讶-yo], [疑问-en], [嘘], [哭泣] | 18 |
| NonVerbalSpeech-38K | [打鼾], [清嗓], [哭泣], [呼吸], [吸鼻子], [笑声], [咳嗽], [喘气], [打哈欠], [叹气] | 10 |
| SMIIP-NV | [笑声], [哭泣], [咳嗽] | 3 |
| Synparaspeech | [叹气], [清嗓子], [笑声], [停顿], [啧声], [倒吸气] | 6 |
| MNV-17 | [打喷嚏], [拍手], [发嘶声], [吹口哨], [清嗓子], [咳嗽], [嘴唇吧唧嘴], [呼气], [呻吟], [喘气], [抽鼻子], [哼唱], [笑声], [鼓掌], [吸气], [轻笑], [叹气] | 17 |

表9：以往研究中使用的非言语标签汇总。

| 聚合类别 | WESR事件标签 |
|---|---|
| 笑声 | [大笑], <正在笑>, [轻笑], [咯咯笑], [人群笑声] |
| 呼喊 | [尖叫], [咆哮], [呼喊], <大喊> |
| 耳语 | <耳语> |
| 歌唱 | <歌唱> |
| 呼吸 | [吸气], [呼气], <喘气>, [叹气], [呼吸] |
| 咳嗽 | [咳嗽], [清嗓子_喉咙] |
| 哭泣 | <哭泣>, [啜泣], [哭] |

表10：非语言事件标签的标签聚合映射。

# F 补充图

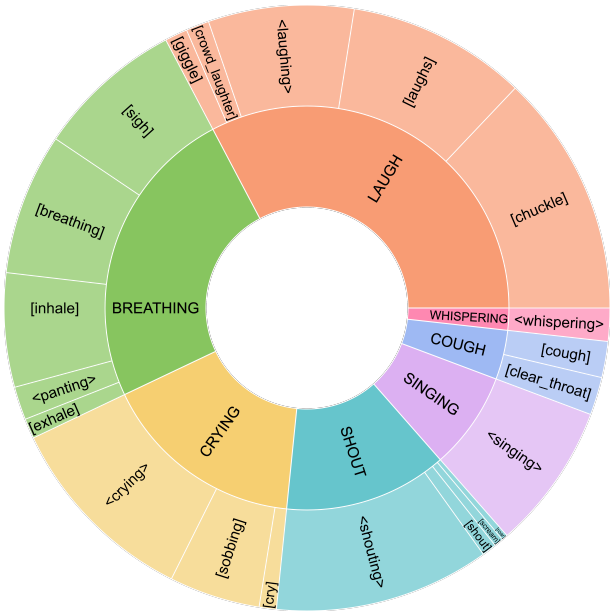


图5：WESR-Train 的标签分布。




图6：人工标注的标注界面过程。