arXiv:2603.11205v1 [eess.AS] 11 Mar 2026

大语言模型能否帮助定位部分伪造语音中的伪造词?

Lin Zhang [1,2], *Thomas Thebaud* [1,2], *ZexinCai* [1,2], *SanjeevKhudanpur* [1,2], *Daniel Povey* [3], *Leibny Paola Garc´ıa-Perera* [1,2], *MatthewWiesner* [1,2], *NicholasAndrews* [1,2]

[1] HLTCOE & [2] CLSP，约翰斯·霍普金斯大学，美国[3] 小米公司，中国北京zlin@ieee.org

摘要

近年来，基于大规模文本训练的大型语言模型（大语言模型）因其在众多任务中的出色表现而备受关注。受此启发，我们探究经过文本训练的大语言模型能否帮助定位部分伪造语音中的伪造词，即语音中只有特定词语经过编辑。我们构建了一个语音大语言模型，通过下一个词元预测来定位伪造词。在 AV-Deepfake1M 和 PartialEdit 上进行的实验与分析表明，该模型经常利用从训练数据中学到的编辑风格模式作为定位伪造词的线索，尤其是我们讨论的这两个数据库中的词级极性替换。尽管这类特定模式在域内场景中能提供有用信息，但如何避免过度依赖这类模式，并提升模型对未见编辑风格的泛化能力，仍是一个悬而未决的问题。

索引词：语音大语言模型，部分伪造，伪造定位

1. 引言

生成式人工智能（GenAI）的进步使高质量合成人类语音及对其进行操控成为可能。这些技术如今能够生成几乎令人类听众和自动检测器都无法与真实语音区分的伪造语音 [1]。当只有少数词语被篡改，而其余部分仍为真实语音时，这种情况称为“部分伪造”[2]，，检测难度会进一步增加 [1]。为支持部分伪造语音研究，已有多个数据集发布：PartialSpoof [2], HAD [3], LlamaPartialSpoof [4],PartialEdit [5], 等。此外，还举办了一些挑战赛：ADD 2022-2023 [6, 7], 和 AV-Deepfake1M [8, 9]。值得注意的是，AV-Deepfake1M [8] 是目前规模最大的一批视听部分伪造数据集之一，而近期发布的 PartialEdit 数据集 [5] 则采用先进的语音填充技术来编辑语音。

在此类部分伪造场景中，定位被篡改的部分至关重要

为定位部分伪造语音中的伪造区域，已有多种方法得到探索 [10]，，这些方法可分为三类：(1) 帧级均匀分段 [3, 11, 12]，对每一帧分别作出判断；(2) 帧级不一致性 [13, 14]，利用相邻帧之间的差异作出判断；(3) 基于边界的定位 [15, 16, 17]，，侧重检测变化点（即真实片段与篡改片段之间的拼接部分）。其中，帧级均匀分段是最简单、最高效的方案。近期一项研究 [18] 表明，采用简单的基于注意力的后端的自监督模型，其效果可与更复杂的伪造定位方案相媲美。

近年来，大语言模型在自然语言处理领域备受关注 [19]，，一旦提出通过基础语音表征（如自监督学习（SSL）表征 [20, 21] 和神经编解码器 [22, 23]）支持语音输入的方案，它们便迅速被应用于语音处理。将这些语音表征投影到与文本共享的大语言模型嵌入空间中，使得大型语音语言模型（语音大语言模型）得以创建 [24, 25]。在生成部分伪造语音时，语音编辑 [26]通常旨在修改词语以改变其原意，即用语义不同的合成词语替换它们。因此，某些特定类型的词语（如反义词）可能更常被修改。此外，一些词语可能会被改成说话人不常使用的词语，或扩展说话人的说话风格。此类修改模式可作为定位伪造词的线索。

同期的两项研究将防御部分伪造语音重新表述为使用语音大语言模型进行问答。Xu 等人 [27] 更侧重于联合建模攻击识别、定位和语义影响。Xue等人 [28]则侧重于联合检测与定位，并利用传统帧级检测器输出的帧级概率。两项研究都展现了语音大语言模型在定位伪造词方面的潜力。然而，它们对大语言模型如何识别部分伪造语音中的编辑区域，所采用的模式分析有限。

因此，本研究旨在探讨的不仅是：“大语言模型能否帮助定位部分伪造语音中的伪造词？”还包括“它们利用了哪些模式来实现这一目标？”本研究聚焦于词级伪造定位，以探究这一问题。我们首先构建一个基线，使用基于自监督学习的模型，将自动语音识别生成的词语与其时间戳对齐，作为参考。随后，我们引入语音大语言模型，通过下一个词元预测来定位伪造词。我们同时考虑无法获取转写文本和可以获取转写文本两种情形。

我们在 AV-Deepfake1M 和 Par- tialEdit 上进行的实验与分析表明，大语言模型依赖于从训练数据中学到的编辑风格模式。例如，这两个数据库都是基于 ChatGPT 编辑以改变含义的。这种编辑通常会通过极性替换产生伪造词，而且这些模式因模态而异。对于仅转录输入，模型主要学习**词汇**线索，例如哪些词**在词语层面**经常被编辑。对于仅音频输入，语音大语言模型主要从**声学**线索中学习，达到**音素级**。尽管此类模式在一定程度上有助于域内伪造词定位，但避免过度依赖数据集特定模式，并提升对未见编辑风格的泛化能力，仍是未来研究的重要方向。

[1]在真实场景中，它被称为“部分伪造” [2].

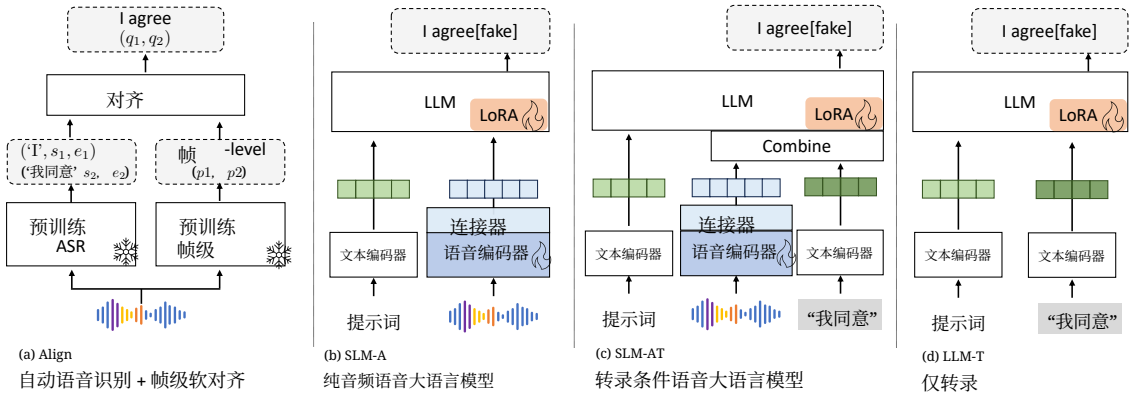


图1：*Fake-word* 模态案例：*(b)* *d* 通过以下方式进行定位：*(a)* 自动语音识别与帧级检测器之间的 *Align*，以及基于大语言模型的方法，适用于以下模态情况：*ree* 仅音频，*(c)* 转录条件音频，以及 *(d)* 仅转录。

2. 从Align到语音大语言模型

如图1所示，本文主要关注两类模型：一种是基于对齐预训练自动语音识别与帧级检测器的基线检测器，另一种是以不同模态作为输入的语音大语言模型。

2.1. 对齐基线

图1(a)展示了我们的基线检测器（Align）。它由一个自动语音识别系统和一个帧级真假检测器组成。帧级检测结果通过自动语音识别生成的时间戳，与词级时间区间对齐。

给定输入 $\boldsymbol{x}$，帧级检测器会输出一系列后验分数 $\boldsymbol{p}_{1:T} = (p_1, \cdots, p_t, \cdots, p_T)$，其中 p_t 表示第 t 帧的预测。同时，预训练的自动语音识别系统会输出一系列带时间戳的词 $\{w_n, s_n, e_n\}$，其中 w_n 是第 n 个词，其开始时间为 s_n，结束时间为 e_n。为了获得词级预测，我们将每个词时间跨度内的帧级分数聚合为每个词的分数：

$$q_n = \text{Agg}(\{p_t | t \in \mathcal{I}_n\}), \qquad \mathcal{I}_n = \{t | t \in [s_n, e_n]\}. \quad (1)$$

其中 Agg(·) 是聚合函数，本研究采用均值聚合：$q_n = \frac{1}{|\mathcal{I}_n|} \sum_{t \in \mathcal{I}_n} p_t$。

2.2. 用于伪造词定位的语音大语言模型

如图1(b-d)所示，我们基于预解码器对齐结构 [25], 构建了一系列语音大语言模型，通过下一个词元预测来定位深度伪造词。它由以下部分组成：(i) 用于生成声学嵌入的语音编码器，(ii) 将音频嵌入投影到大语言模型文本特征空间的连接器模块，以及 (iii) 大语言模型。

给定语音和带有可选转写文本的提示词，大语言模型将执行下一个词元预测，在任何被检测为伪造的词之后附加"[fake]"。通过微调 Whisper [29]，采用下一个词元预测进行伪造词定位的类似策略已在同期工作中被证明有效。

用于伪造词定位的语音大语言模型有两个目的：(i) 转写语音，以及 (ii) 定位伪造词。据此，我们考虑两种情形：

1. 没有转录文本，仅提供音频。此时，大语言模型需要转写音频并定位伪造词。2. 有转录文本。此时，任务会变得更简单，大语言模型只需利用语音作为额外证据。

因此，本研究讨论了三种输入模态条件，如图1(c–d)所示：(b) 场景1中以仅音频作为输入的 SLM-A；(c) 场景2中有转录文本时使用的音频+转写（SLM-AT）。此外，还将仅转录（LLM-T）作为参考，以考察大语言模型如何从转录文本中学习。

3. 实验设置

3.1. 数据库

本文讨论了两个规模不同的数据库：

PartialEdit（PE） [5] 是近期发布的部分伪造数据库，由语音填充或编辑生成。PartialEdit 基于 VCTK [30], 构建，其中包含 110 位说话人，每人有 400 条朗读句子。这些句子选自新闻报纸、彩虹段落和引导段落。因此，我们可以获取其真实转录文本。与 AV1M 一样，我们使用 ChatGPT [31]修改这些转写文本以改变句意，随后使用一组语音合成系统对语音进行编辑。受许可限制，本研究仅使用 PartialEdit 中的 E1（VoiceCraft[32]）和 E2（SSR[33]）子集。

AV-Deepfake1M (AV1M)[2] [8]基于 VoxCeleb2 [34] 构建。VoxCeleb2 是一个从 YouTube 收集的名人语音数据集，最初用于训练说话人验证系统，主要包含自发的公开采访和演讲。生成部分伪造媒体时，我们首先使用 Whisper [35], 获取转录文本，再使用 ChatGPT [31]生成目标操纵内容，以反向修改原意。随后，使用 YourTTS [36]和 VITS [37] 系统生成与修改后词语对应的音频，同时保持其余部分不变。我们从 AV-Deepfake1M视频中提取音轨。AV1M 包含四种样本类型：音频修改、视觉修改、音频和视觉均修改，以及真实。由于我们专注于音频分析，因此仅使用音频修改和真实子集。此外，由于评估集的元数据尚未发布，我们仅使用提供的训练集和验证子集。

[2]我们没有使用最新版本的 AV-Deepfake1M++ [9]，因为部分验证集元数据不可用。

表1：使用*Align*模型的跨域*WordF1*（%）。

训练	Test		
	PE-dev	PE-eval	AV1M-dev
PE-train	54.28	53.75	12.03
AV1M-train	0.04	0.04	41.48
词错误率 (%)	2.20	1.94	11.99†

† 请注意，AV1M 上的词错误率仅供参考，因为其转录参考文本由 Whisper 生成，而非通过人工标注生成 [8]。

表 2：大语言模型在不同输入模态下的域内词错误率和 *WordF1*（%）。*(a) PE*：在 *PE-train* 上训练并在 *PE-eval* 上评估的模型。*(b) AV1M*：在 *AV1M-train* 上训练并在 *AV1M-dev* 上评估的模型。

模态	(a) PE		(b) AV1M	
	词错误率 ↓	词F1 ↑	词错误率 ↓	词F1 ↑
SLM-A	62.89	54.08	11.63	94.68
LLM-T	0.07	83.63	0.00	84.80
SLM-AT	0.02	90.79	0.01	97.51

3.2. 实验设置

对于 Align 模型，我们需要：(1) 一个自动语音识别模型，用于生成带有词级时间戳的转写文本；(2) 一个定位模型，用于生成伪造区域的时间戳。对于自动语音识别，我们使用预训练的 parakeet-tdt-0.6b-v2[3] 获取词级时间戳。对于帧级伪造定位，鉴于 SSL-MHFA [38]在轻量化设计下表现相对更好，我们采用该模型 [18] ，其配置与 [38]中的相同。

对于语音大语言模型，我们基本沿用了 [39] 使用其开源代码[4]时的配置。具体而言，我们使用 WavLM-base-plus[5] 作为语音编码器，并使用 TinyLlama[6] 作为大语言模型组件，因为它仅有 11 亿个参数，却具备出色的能力。WavLM 与 Llama 的这种组合在自动语音识别任务中表现相对更好 [40]。根据我们的初步比较，我们采用两层常规连接器。微调语音编码器时，学习率设为 2e-06。大语言模型和连接器的学习率设为 0.0001。LoRA 的秩设为=8 ，alpha 设为=16。

3.3. 评估指标

以往大多数关于伪造词定位的工作都强调细粒度时间精度，评估指标也相应地针对细粒度时间戳而设计。因此，我们引入词级指标WordF1，用于本研究中评估词级伪造定位。

这里，我们采用词级 F1（WordF1）来衡量模型检测伪造词的能力，并将伪造词视为正类。我们使用 JIWER[7]，根据莱文斯坦距离对齐参考文本和假设文本中的词。在实际攻击场景中，攻击者通常只修改极少数词语，以改变语音的语义，同时保持整体自然度。因此，伪造词的比例远低于真实词。使用 WordF1 并将

[3]https://huggingface.co/nvidia/parakeet-tdt-0 2 4 6.b-v
https://github.com/thomasthebaud/speechLLM [5]
https://huggingface.co/microsoft/wavlm-base-plus [6]
https://huggingface.co/TinyLlama/TinyLlama-1 0 7 1
1.B-Chat-v. https://jitsi.github.io/jiwer

表3：使用*SLM-AT*的跨域*WordF1*（%）。

训练	Test		
	PE-dev	PE-eval	AV1M-dev
PE-train	91.10	90.79	32.26
AV1M-train	0.04	0.09	97.51

伪造词视为正类，可以更有效地评估模型在这一少数类上的表现，并有助于避免因将大多数词预测为占多数的真实类而人为获得较高的 F1 分数。

请注意，WordF1 是基于与自动语音识别结果对齐的词语计算的，前提是这些词语已被准确识别，或者替换错误已与参考文本对齐。因此，WordF1 的可靠性与自动语音识别的性能密切相关。为此，我们还报告词错误率（WER）。如果词错误率过高，由于参考文本与假设文本之间未对齐，词级定位指标可能会覆盖较少的词语。

4. 结果

4.1. 使用 Align 模型进行伪词定位

我们首先在表1中为 Align 构建跨域实验。这些结果作为参考，用以评估基于语音大语言模型的模型是否取得了合理的性能。

从表1的域内实验来看，PE 取得了更低的词错误率和更高的 WordF1。从表1的跨域结果来看，当使用在 AV1M-train 上训练并在 PE-dev 和 PE-eval 上评估的模型时，我们观察到极低的 WordF1（0.04）。一个可能的原因是 AV1M 中的假词比例较低，这使模型偏向于预测“真实”，从而导致在域外条件下对 PE 出现大量漏检。这凸显了 Align 模型的泛化能力有限。

4.2. 大语言模型能否帮助定位部分伪造语音中的伪造词？——不同模态下的分析。

我们按照第2.2节所述，在不同输入模态下使用大语言模型来探讨这一问题：仅音频（SLM-A）、仅转录（LLM-T），或两种模态（SLM-AT）。PE 和 AV1M 的域内结果分别展示在表2(a)和(b)中。

我们首先考察第一行（SLM-A）。该模型仅接受音频输入，与第 4.1 节中 Align基线的输入条件相同。在 PE-eval 上，SLM-A 难以生成准确的转写文本（词错误率 = 62.89%）。但在采用其他对齐转写文本的情况下，其 WordF1=54.08% 与 Align 方法相当，这表明语音大语言模型能够在一定程度上定位伪造词。在 AV1M-dev 上，SLM-A 达到了与 Align 相近的词错误率（11.63%），但 WordF1 高得多（94.68%），这说明当其转写质量足够高时，语音大语言模型能够有效定位伪造词。这回答了我们的问题：语音大语言模型可以在域内场景下定位部分伪造语音中的伪造词，但在某些设置中可能会受到转写文本质量的限制。如何在使用语音大语言模型定位伪造词的同时改进仅音频转写，留待未来研究。

因此，我们考察了第 2.2 小节提出的第二种情形：已有转写文本，语音大语言模型只需定位伪造词。在提供转写文本后，模型（LLM-T 和 SLM-AT）基本保留了输入文本（词错误率接近于零），表明文本几乎未被改动，也几乎没有产生幻觉。尤为值得注意的是，将音频与转写文本结合起来

表4： 不同输入模态下训练数据和测试数据中排名前*10*的伪造词汇及音素线索。两个列表均按频率排序。列标识符以括号标出（列标识符）。

数据集	训练		AV1M-dev (SLM-A) 上的预测模式		AV1M-dev (LLM-T) 上的预测模式	
	(1) 词语	（2）音素	（3）单词	（4）音素	（5）单词	（6）音素
PE-train	忽略，一个，坏的，糟糕的，ig-未保存，被拒绝，迷失，失去，容易的，一个	AH，N，T，D，IH，S，R，L，K，EH，IY，ER，P，M，EY	银，旅程，去，经理，蓝色，吸收，在，是，你，数据	AH, T, IH, N, S, L、R、ER、D、IY、V, P, K, EH, Z	糟糕的，不喜欢，坏的，讨厌，无聊的，总是，失望了，做，普通，最差	AH，T，IH，S，L，N, D, R, K, EH, IY, B, P, ER, M
AV1M-train	糟糕的，do，not，不喜欢，无聊的，非常，讨厌，did，绝对，t	AH, T, IH, N, L, D, S, R, IY, K, EH, M, B, ER, P	糟糕的，do，not，不喜欢，坏的，做了，讨厌，无聊的，非常，easy	AH, T, IH, N, D, L, S, R, IY, K, EH, B, F, M, ER	糟糕的，do，不喜欢，bad，讨厌，不，做，无聊的，非常，t	AH, IH, T, D, L, N, S, R, IY, K, EH, B, ER, M, F

并使用 SLM-AT，始终比仅使用 LLM-T 处理转写文本取得更高的 WordF1，这表明音频为定位伪造词提供了互补线索（在 AV1M 上，WordF1 从 84.80% 提高到 97.51%；在 PE-eval 上，从 83.63% 提高到 90.79%）。

我们还使用 SLM-AT 评估了跨域表现，如表3所示。在 AV1M 上训练并在 PE 上测试时，模型的 WordF1 极低，且倾向于将所有词都标记为真实。第 4.1 节中的 Align 基线也呈现出类似模式，这表明泛化能力差是两种方法都难以应对的普遍挑战。

其他研究也报告了在使用基于语音大语言模型的问答进行伪造词定位时存在泛化问题 [27]。这引出了一个问题：为什么语音大语言模型在跨域场景下性能会下降?

另一个值得注意的观察是，即使没有音频，在域内设置中，模型仅凭仅转录文本输入也能达到 83.63%。大语言模型会利用哪些模式，在不同模态下定位伪造词?

4.3. 大语言模型利用哪些模式来定位伪造词?

为了探究跨域条件下性能为何下降，以及大语言模型如何在不同模态下进行定位，我们对训练数据和模型预测中伪造词的频率及这些伪造词中的音素分布进行了初步分析。表4按频率列出了排名前10的伪造词和音素。分析主要针对难度较高的 AV1M-dev。第一行对应使用 PE-train（跨域）训练的模型，第二行对应使用 AV1M-train（域内）训练的模型。对于每个训练集，我们首先列出该训练数据中常见的词和音素（第1–2列）。随后，我们报告在 AV1M-dev 上评估的特定模态模型的统计数据：第3–4列为 SLM-A（仅音频）的预测结果，第5–6列为 LLM-T（仅转录）的预测结果。

从表4第1–2列可以看出，两种训练数据中呈现出相似的编辑模式：许多词语被改成带有负面情感的表达（例如“坏的”“糟糕的”）。这可以理解，因为正如第3.1节所述，原始语音大多面向公众，且倾向于表达积极情感。当提示 [8], ChatGPT“将其含义反转为相反的方向”时，它通常会用反义词替换原词，经常将积极词语换成消极对应词，从而改变话语的整体含义。仅以转写文本作为输入的 LLM-T 很容易学会这些编辑模式，这一点也反映在最后的第5–6列预测结果中。

从 SLM-A 的预测中也能观察到相似但有所不同的模式，如第 3–4 列所示。当模型以仅音频作为输入时，它会学习操纵的**语音** 线索。在频繁预测的词语中，“银”的出现频率最高，尽管它并未出现在训练数据中排名前 10 的词语里。这可能是因为模型检测到其音素序列 [S IH L V ER] 主要由常见的训练音素构成（/V/ 除外）。这表明 SLM-A 倾向于依靠音素级声学证据来定位伪造词。

尽管这些编辑模式能够在一定程度上帮助大语言模型在域内定位伪造词，因为模型可以据此捕捉原句被修改的方式，但过度依赖这些模式可能导致模型忽视其他伪造痕迹，使其倾向于定位特定模式（如反义词或负面词语），而非真正检测深度伪造词。这或许可以解释模型在跨域条件下性能下降的原因。在真实场景中，编辑方式可能更加多样且隐晦：除了反转含义，攻击者还可能修改命名实体。因此，过度依赖训练数据中的特定模式会降低模型在实际环境中的表现。在使用大语言模型时，避免过度依赖数据集特有模式并提升对未见编辑风格的泛化能力，仍是未来研究的重要方向。

5. 结论

本研究探讨了“大语言模型能否帮助定位部分伪造语音中的伪造词？”为此，我们通过下一个词元预测构建了用于伪造词定位的语音大语言模型。我们考虑了两种场景：转录文本不可用时和可用时。结果表明，语音大语言模型能够在域内场景下定位部分伪造语音中的伪造词，但在跨域场景下性能下降。

为了理解“为何语音大语言模型在跨域场景下性能下降？”以及“大语言模型如何帮助定位伪造词？”，我们初步分析了训练数据和模型预测中词汇相关词语与声学相关音素的模式。以仅转录作为输入时，大语言模型通过从训练数据中学习编辑模式来帮助在域内场景下定位伪造词，这些模式通常是反义词替换。而当以仅音频作为输入时，语音大语言模型尽管此类特定模式可能有助于定位被编辑的词语，过度依赖训练数据中的特定模式因此会降低其在真实环境中的性能。在利用大语言模型时，避免过度依赖数据集特定模式，并提升对未见编辑风格的泛化或适应能力，仍是未来工作的重要方向。

6. 致谢

作者感谢清华大学的 Kai Li 提供讨论和宝贵建议。本研究得到国家情报总监办公室（ODNI）和情报高级研究计划局（IARPA）的支持，资助来自 ARTS计划，合同编号为 D2023-2308110001。

7. 生成式人工智能披露

生成式人工智能工具仅用于编辑和润色由人工撰写的初稿。所有经人工智能辅助生成的文本均已由作者在提交前审核。

8. 参考文献

[1] A. Alali 和 G. Theodorakopoulos，“利用深度伪造音频在真实世界中实施伪造语音攻击”，《网络安全与隐私期刊》，第 5 卷，第 1 期，第 6 页，2025 年。[2] L. Zhang、X. Wang、E. Cooper、J. Yamagishi、J. Patino 和 N. Evans，“检测部分伪造音频的初步研究”，载于*Interspeech* 会议论文集，2021 年，第 4264–4268 页。[3] J. Yi、Y. Bai、J. Tao、H. Ma、Z. Tian、C. Wang、T. Wang 和 R. Fu，“Half-Truth：部分伪造音频检测数据集”，载于*Interspeech 2021* 会议论文集，2021 年，第 1654–1658 页。[4] H.-T. Luong、H. Li、L. Zhang、K. A. Lee 和 E. S. Chng，“Llama-partialspoof：一个由大语言模型驱动、模拟虚假信息生成的伪造语音数据集”，载于*ICASSP* 会议论文集。IEEE，2025 年，第 1–5 页。[5] Y. Zhang、B. Tian、L. Zhang 和 Z. Duan，“PartialEdit：神经语音编辑时代的部分深度伪造识别”，载于*Interspeech* 会议论文集，2025 年，第 5353–5357 页。[6] J. Yi、R. Fu、J. Tao、S. Nie、H. Ma、C. Wang、T. Wang、Z. Tian、Y. Bai、C. Fan、S. Liang、S. Wang、S. Zhang、X. Yan、L. Xu、Z. Wen 和 H. Li，“ADD 2022：首届音频深度合成检测挑战赛”，载于*ICASSP* 会议论文集，2022 年，第 9216–9220 页。[7] J. Yi、J. Tao、R. Fu、X. Yan、C. Wang、T. Wang、C. Y. Zhang、X. Zhang、Y. Zhao、Y. Ren 等，“ADD 2023：第二届音频深度伪造检测挑战赛”，载于*IJCAI DADA* 会议论文集，2023 年。[8] Z. Cai、S. Ghosh、A. P. Adatia、M. Hayat、A. Dhall、T. Gedeon 和 K. Stefanov，“AV-Deepfake1M：一个由大语言模型驱动的大规模视听深度伪造数据集”，载于*ACM MM* 会议论文集，2024 年。[9] Z. Cai、K. Kuckreja、S. Ghosh、A. Chuchra、M. H. Khan、U. Tariq、T. Gedeon 和 A. Dhall，“AV-Deepfake1M++：一个具有真实场景扰动的大规模视听深度伪造基准”，载于*ACM MM* 会议论文集，2025 年，第 13 686–13 691 页。[10] J. He、J. Yi、J. Tao、S. Zeng 和 H. Gu，“部分深度伪造音频中篡改区域定位综述”，*arXiv* 预印本 *arXiv:2506.14396*，2025 年。[11] L. Zhang、X. Wang、E. Cooper、N. Evans 和 J. Yamagishi，“partialspoof 数据库及其反制措施：检测嵌入语句中的短伪造语音片段”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 31 卷，第 813–825 页，2023 年。[12] Y. Xie、H. Cheng、Y. Wang 和 L. Ye，“一种基于嵌入表示的高效临时深度伪造定位方法，用于部分伪造音频检测”，载于*ICASSP* 会议论文集，2024 年，第 966–970 页。[13] J. Wu、W. Lu、X. Luo、R. Yang、Q. Wang 和 X. Cao，“用于音频时序伪造检测与定位的由粗到细提案细化框架”，载于*ACM MM* 会议论文集，2024 年，第 7395–7403 页。[14] M. Li、X.-P. Zhang 和 L. Zhao，“用于部分深度伪造语音检测的帧级时序差异学习”，《*IEEE* 信号处理快报》，2025 年。[15] J. Zhong、B. Li 和 J. Yi，“利用边界感知注意力机制增强部分伪造音频定位”，载于*Interspeech* 会议论文集，2024 年，第 4838–4842 页。[16] H. Wu、H.-C. Kuo、N. Zheng、K.-H. Hung、H.-Y. Lee、Y. Tsao、H.- M. Wang 和 H. Meng，“基于自注意力的伪造片段发现方法，用于部分伪造音频检测”，载于*ICASSP* 会议论文集，2022 年，第 9236–9240 页。[17] Z. Ge、X. Xu、H. Guo、Z. Yang 和 B. Schuller，“GNCL：一种结合一致性损失的图神经网络，用于片段级伪造语音检测”，载于*ICASSP* 会议论文集，2025 年，第 1–5 页。

[18] L. Zhang, J. Rohdin, X. Wang, J. Peng, T. Liu, Y. Zhang, H.-T. Lu- ong, S. Wang, C. Liang, A. Silnova 等，“WeDefense：一种防御伪造音频的工具包”，arXiv 预印本 arXiv:2601 15240，2026 年。.[19] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong 等，“大型语言模型综述”，arXiv 预印本 arXiv:2303.18223，第 1 卷，第 2 期，第 1–124 页，2023 年。[20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao 等，“WavLM：面向全栈语音处理的大规模自监督预训练”，IEEE信号处理精选主题期刊，第 16 卷，第 6 期，第 1505–1518 页，2022 年。[21] A. Baevski, Y. Zhou, A. Mohamed 和 M. Auli，“wav2vec 2.0：语音表征自监督学习框架”，载于 NeurIPS 论文集，第 33 卷，2020 年，第 12 449–12 460 页。[22] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar 和 K. Kumar，“采用改进 rvqgan 的高保真音频压缩”，载于 NeurIPS 论文集，第 36 卷，2023 年，第 27 980–27 993 页。[23] A. D´efossez, J. Copet, G. Synnaeve 和 Y. Adi，“高保真神经音频压缩”，机器学习研究汇刊，2023 年。[24] S. Arora, K.-W. Chang, C.-M. Chien, Y. Peng, H. Wu, Y. Adi, E. Dupoux, H. yi Lee, K. Livescu 和 S. Watanabe，“口语语言模型研究概况：全面综述”，机器学习研究汇刊，2025 年。[25] J. Peng, Y. Wang, B. Li, Y. Guo, H. Wang, Y. Fang, Y. Xi, H. Li, X. Li, K. Zhang, S. Wang 和 K. Yu，“面向理解的语音大语言模型综述”，IEEE信号处理精选主题期刊，第 20 卷，第 1 期，第 2–31 页，2026 年。[26] T. K¨assmann, Y. Liu 和 D. Liu，“语音编辑综述”，arXiv 预印本 arXiv:2407.17172，2024 年。[27] X. Xu, Y. Ren, L. Liu, W. Wu, B. Li, C. Lu, S. Wang 和 C. Zhang，“Holiantispoof：用于全面语音反欺骗的音频大语言模型”，arXiv 预印本 arXiv:2602.04535，2026 年。[28] J. Xue, Y. Chai, Y. Ren, J. He, Z. Tang, Z. Yi, Y. Huang, Y. Xie 和 Y. Chen，“通过先验增强音频大语言模型统一语音编辑检测与内容定位”，arXiv 预印本 arXiv:2601.21463，2026 年。[29] H. M. Tran, X. Wang, W. Ge, X. Liu 和 J. Yamagishi，“使用微调后的 Whisper 通过下一个词元预测检测深度伪造词语”，arXiv 预印本 arXiv:2602.22658，2026 年。[30] C. Veaux, J. Yamagishi 和 K. MacDonald，“CSTR VCTK 语料库：供 CSTR语音克隆工具包使用的英语多说话人语料库”，语音技术研究中心（CSTR），2016 年。[31] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Ale- man, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat 等，“GPT-4 技术报告”，arXiv 预印本 arXiv:2303.08774，2023 年。[32] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed 和 D. Harwath，“VoiceCraft：面向真实场景的零样本语音编辑与文本转语音”，载于 ACL 论文集，2024 年，第 12 442–12 462 页。[33] H. Wang, M. Yu, J. Hai, C. Chen, Y. Hu, R. Chen, N. Dehak 和 D. Yu，“SSR-speech：迈向稳定、安全、鲁棒的零样本文本语音编辑与合成”，载于 ICASSP 论文集。IEEE，2025 年，第 1–5 页。[34] J. S. Chung, A. Nagrani 和 A. Zisserman，“VoxCeleb2：深度说话人识别”，载于 Interspeech 论文集，2018 年，第 1086–1090 页。[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey 和 I. Sutskever，“通过大规模弱监督实现鲁棒语音识别”，载于 ICML 论文集，第 202 卷。PMLR，2023 年 7 月 23–29 日，第 28 492–28 518 页。[36] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. G¨olge 和 M. A. Ponti，“YourTTS：迈向人人可用的零样本多说话人文本转语音与零样本语音转换”，载于 ICML 论文集，2022 年，第 2709–2720 页。

[37] J. Kim, J. Kong 和 J. Son，“采用对抗学习的条件变分自编码器端到端文本转语音”，载于 *ICML* 论文集，2021 年，第 5530–5540 页。[38] J. Peng, O. Plchot, T. Stafylakis, L. Mosner, L. Burget 和 J. Cer- nocky，“一种支持高效微调 Transformer模型的基于注意力的后端，用于说话人验证”，载于 *SLT* 论文集，2023 年，第 555–562 页。[39] T. Thebaud, Y.-J. Lu, M. Wiesner, P. Viechnicki 和 N. De- hak，“利用说话人特征增强对话标注：借助冻结的大语言模型”，载于 *ASRU* 论文集，2025 年。[40] A. Fortier, T. Thebaud, J. Villalba, N. Dehak 和 P. Cardinal，“针对语音语言模型的后门攻击”，*arXiv* 预印本 *arXiv:2510.01157*，2025 年。