arXiv:2611.29085v1 [cs.SD] 21 Nov 2026

基于边界回归的欺骗定位

Yixuan Xiao[1]* *Xin Wang*[2]†, *Yassine El Kheir*[3], *Feidi Kallel*[3], *Arnab Das*[3], *Tim Polzehl*[3], *Sebastian M¨oller*[4], *Ngoc Thang Vu*[1]

[1]斯图加特大学，德国 [2]日本国立信息学研究所，日本[3]德国人工智能研究中心（DFKI），德国[4]柏林工业大学，德国

摘要

部分伪造音频在真实语音中嵌入伪造片段，而欺骗定位的任务就是定位这些片段。大多数反欺骗对策（CM）在帧级运行。通常，它们包括一个将音频表示为帧的前端，以及一个将每帧分类为真实语音或欺骗的后端。这种帧级检测将参考值、预测结果和评估都绑定到预定义的时间分辨率：包含过渡的帧只能接收单一标签，边界不能位于帧内部，并且评估也取决于该分辨率。为解决这些问题，我们将欺骗定位视为回归问题。我们提出了一种单阶段边界回归器，输出具有连续值时间戳的（起点，终点、分数）元组，并基于精确的参考区间进行训练。结果表明，所提出的反欺骗模型优于以往的帧级和基于边界的反欺骗对策，并且具有更好的泛化能力。

索引词— 欺骗定位、边界回归、反欺骗

1. 引言

在部分伪造（PS）场景中，攻击者不会合成整段话语，而是用伪造语音替换真实话语中的几个词，从而改变其含义，同时大部分音频仍为真实语音 [1, 2, 3]。对于此类音频，仅在话语层面做出判断并不总是足够的。在取证分析等应用中，人们可能需要知道话语中的哪一部分被编辑过。定位部分伪造音频中的伪造区域这一任务称为欺骗定位。

大多数用于欺骗定位的反欺骗对策（CM）都将该任务视为分类问题。典型的反欺骗模型利用自监督学习（SSL）模型作为前端 [4], ，将音频表示为一系列称为帧的特征向量，并使用后端将每一帧分类为真实语音或欺骗。请注意，帧数由 SSL 卷积层的步幅决定；在 wav2vec2 风格前端 [5, 6, 7, 8], 中，连续帧相隔 20 毫秒，我们将这一距离称为帧分辨率。

然而，伪造词的起点和结束时间不一定是 20 ms 的整数倍，而帧级CM只能以 20 ms 帧为单位进行观察和预测。这种不匹配会导致三个具体问题。**第**一，参考标签必须量化到帧分辨率，这会在训练中引入不可避免的量化误差。具体而言，创建部分伪造音频时，伪造区间 $[t_s, t_e]$ 是已知的精确值，且独立于反欺骗模型的时间分辨率。但要为每一帧分配一个标签，

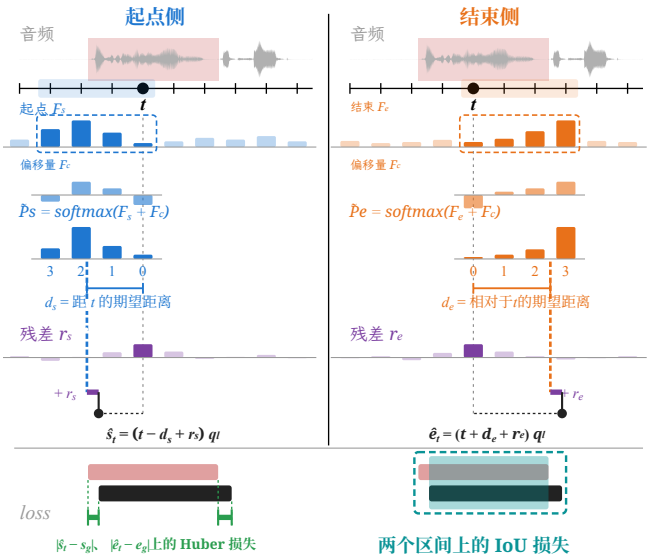

图 1。单个层级中某一时刻的边界解码。为该层级的所有时刻预测起始和结束分数；某一时刻 t 会在每一侧将自身偏移量加到最近的分数上，将其转换为概率，并取期望距离加上一个较小的残差作为边界。使用 IoU 损失训练区间，并使用 Huber 损失训练各边界与参考值之间的间隔。

完全被伪造的帧和只有一小部分被伪造的帧会被赋予相同的“欺骗”标签，尽管它们的“欺骗程度”不同。**第**二，模型输出的精度受分辨率限制，因为帧级反欺骗模型无法预测帧内的边界。**第**三，评估也与分辨率相关：基于点的等错误率会随着帧评分所采用的分辨率而显著变化，因此结果通常会以不同分辨率报告，使得比较不那么直接。虽然使用基于范围的等错误率 [9], 可以缓解评估方面的问题，但其理论基础尚未得到充分论证。[1] 更重要的是，它并未解决前两个问题。

为解决上述问题，我们通过相对边界建模将欺骗定位视为回归问题。受 TriDet [10],启发，我们提出了一种反欺骗模型，该模型以一段话语作为输入，并输出一个元组列表（s，e，c），其中 s，e，$c \in \mathbb{R}$；s 和 e 定义了推测欺骗区间的时间范围，而 c是其欺骗分数。

*通讯作者：yixuan.xiao@ims.uni-stuttgart.de

†感谢JST PRESTO资助项目（JPMJPR23P9）提供资金支持。

[1]见附录A。

我们的**贡献**[2] 包括：(1) 一种单阶段反欺骗模型，可回归精确的欺骗边界，在训练、预测或评估过程中均不使用量化标签。(2) 在 PartialSpoof 上，该反欺骗模型取得了最低的基于点的等错误率和基于范围的等错误率；与其他最先进方法相比，它对 LlamaPartialSpoof 的泛化能力更强。(3) 该反欺骗模型揭示了已发布的 PartialEdit 参考值中存在时间错位问题，并使用互相关发布了修正版。

2. 方法

所提出的反欺骗模型保留了帧级CM的自监督学习前端：一个 300M 的 XLS-R 前端将话语表示为由 1024 维帧构成的序列。其 Transformer模块经过微调，而卷积特征提取器则保持冻结。如图 2 所示，从这些特征中预测的内容有三方面变化：(1) 将帧序列扩展为特征金字塔，使每个位置都能看到输入中不同范围的内容（2.1）；(2) 不再为每帧预测一个标签，而是在每个层级的每个位置回归该位置周围伪造区间的精确起始时间和结束时间，并预测该区间的欺骗分数（2.2）；(3) 话语级头充当门控，决定是否应为某个话语输出任何伪造区间（2.2）。

我们将所得帧序列在金字塔层级 l 上的一个位置称为瞬时点 $t \in \{0, 1, \ldots, T_l - 1\}$，其中 T_l 是该层级的瞬时点数量。对于每个金字塔层级 l上的每个瞬时点 t，反欺骗模型都会生成一个候选 $(\hat{s}_t, \hat{e}_t, \hat{c}_t)$。$\hat{c}$ 是由起始时间 $\hat{s}_t$ 和结束时间 $\hat{e}_t$定义的区间的欺骗分数：

$$\begin{aligned} \hat{s}_t &= \Delta\big[(t - d_s + r_s)\, q_l\big] + \delta, \\ \hat{e}_t &= \Delta\big[(t + d_e + r_e)\, q_l\big] + \delta, \end{aligned} \tag{1}$$

其中， Δ 是帧分辨率， q_l 是第 l 层的步幅（第2.1节）。尽管 $\hat{s}_t$ 和 $\hat{e}_t$ 的值主要由 $\Delta\, t\, q_l$决定，但额外的偏移项 d_s、 $d_e \in \mathbb{R}_+$ 和 r_s、 $r_e \in \mathbb{R}$ 有助于更精确地预测起始时间和结束时间。其中，d_s、 d_e 是通过概率建模得到的从时间 $\Delta\, t\, q_l$ 到实际起始时间或结束时间的距离（第2.2节）； r_s、 $r_e \in (-0.5,\ 0.5)$ 是有界残差，用于将每个边界进一步细化至最多半个步幅（第2.2节）。最后， δ 是一个常数，属于实现细节，其作用类似于STFT中的左侧填充。

2.1. 多粒度建模

本部分遵循基于XLS-R特征构建的传统帧级CM：使用两个卷积核大小为3的一维卷积模块，将前端输出投影到 $C = 256$ 个通道，然后按照文中所述将序列转换为特征金字塔

多分辨率 CM [2]，可捕获由细到粗的特征。如图 2 所示，金字塔在 20 至 640 毫秒范围内包含六个层级，每个层级都由前一个层级通过步幅为 2 的最大池化构建而成，因此层级 l 的步幅为 $q_l = 2^{l-1}$ 帧。

随后，每个层级都由文献 [10]提出的一个 SGP 模块进行处理。其核心函数定义如下：

$$\begin{aligned} F_{\mathrm{SGP}}(\mathbf{X}) &= \phi(\bar{\mathbf{x}}) \odot \mathrm{FC}(\mathbf{X}) \\ &\quad + \psi(\mathbf{X}) \odot \big(\mathrm{Conv}_w(\mathbf{X}) + \mathrm{Conv}_{kw}(\mathbf{X})\big) + \mathbf{X}, \end{aligned} \tag{2}$$

其中，$\mathbf{X} \in \mathbb{R}^{C \times T_l}$ 是该层包含 T_l 个时间点的特征序列，而 $\bar{\mathbf{x}} \in \mathbb{R}^C$ 是其在这些时间点上的均值。 $\phi(\bar{\mathbf{x}})$ 只根据均值计算一次，并

[2]https://github.com/XIAOYixuan/boundary_detection/

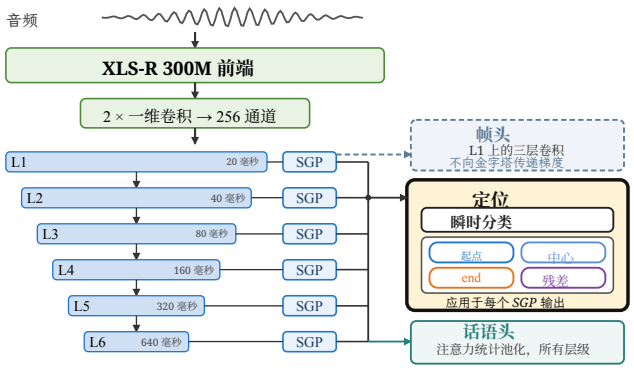

图 2。架构概览。

应用于每个时间点，因此每个时间点都根据话语级上下文进行加权。 ψ、Conv_w 和 Conv_{kw} 是采用两种窗口大小 $w = 1$ 和 $kw = 3$的逐通道卷积，用于建模每个时间点的局部上下文。在每个 SGP 模块中，输入会进行层归一化，公式 (2) 的输出接着经过组归一化和一个前馈网络，并采用第二个残差连接，类似于 TriDet [10]。随后将得到的特征输入不同的任务头以执行不同任务。

2.2. 预测任务

将话语的参考值定义为一组以秒为单位的伪造区间 $G = \{ (s_g,\ e_g) \}$，与发布的数据完全一致。真实话语具有 $|G| = 0$。下文所述拟议反欺骗模型的三个任务均针对这些精确区间进行训练；参考值不会被量化。仅为与其他帧级反欺骗模型进行比较，额外加入了一个辅助任务——帧级预测。这是唯一使用量化标签的任务，并且梯度在预测头处被 截断，因此潜在的量化误差永远不会传递到拟议的反欺骗模型中。

采用中心采样的瞬时分类。该任务检测某一瞬时是否接近参考伪造区间的中心。该任务的预测头是一个三层卷积模块：前两层的卷积核大小为 3、 $C = 256$个通道，并使用 ReLU；最后一层映射为一个分数。

令 $\mathbf{H}^l \in \mathbb{R}^{C \times T_l}$ 表示 SGP 模块在层级 l的输出，则对于六个层级，共有六个输出 $\mathbf{H}^1$、.....、$\mathbf{H}^6$。预测头分别处理每个 $\mathbf{H}^l$，并将同一个预测头应用于全部六个层级。其输出长度 T_l 与输入相同，因此每个瞬时都有一个欺骗分数 $\hat{c}_t$ 。同一个 $\hat{c}_t$ 被用作在 t处预测的区间的欺骗分数。通过中心采样 [11]，（一种也用于 [12, 10]的方法），将每个瞬时 t 标记为正类（接近中心）或负类（不接近中心）；即，中心采样将每个正类瞬时分配给一个参考伪造区间 g。预测头使用二元焦点损失 $\mathcal{F}_t$ [13]进行训练，具体如下：

$$\mathcal{L}_{\mathrm{cls}} = \frac{1}{|\mathcal{P}|} \sum_{t \in \mathcal{P}} \mathrm{IoU}_t\, \mathcal{F}_t + \frac{1}{|\mathcal{N}|} \sum_{t \in \mathcal{N}} \mathcal{F}_t, \tag{3}$$

其中， $\mathcal{P}$ 和 $\mathcal{N}$ 分别是一个批次中的正类和负类时刻集合。IoU_t 是 t 处预测的区间与分配给 t的参考伪造区间之间的交并比。它对正类样本进行加权，使得 $\hat{c}_t$ 能够反映区间的定位效果 [10]。

边界回归。 该任务针对每个正类时刻 t，预测伪造区间的精确起点和终点。根据公式 (1)，这需要距离 d_s, d_e 以及残差 r_s, r_e。如图所示

图 1 中，它们由四个任务头进行预测：起点头、终点头和偏移头共同给出两个距离，而残差头给出两个残差。每个任务头都采用相同的三层卷积结构，并接收与时刻分类头相同的 SGP 模块输出。

起点头和终点头的输入与时刻分类头相同：每个任务头都以 $\mathbf{H}^l$ 为输入，并在全部六个层级之间共享。距离预测遵循 TriDet [10]。起点头和终点头针对每个时刻输出一个分数，$F_s \in \mathbb{R}^{T_l}$和 $F_e \in \mathbb{R}^{T_l}$，分别表示该时刻是起点或终点的可能性。

偏移头的输入也与时刻分类头相同。偏移头输出 $F_c \in \mathbb{R}^{T_l \times 2 \times (B+1)}$。对于时刻 t，它会调整该时刻以及其两侧 $B = 16$ 邻居的分数 t 。以起始时间预测为例：对于t，该时刻或其邻居是起点的概率 $P_s(b)$为 $\mathrm{softmax}_b\big(F_s[t-b] + F_c[t,0,b]\big)$，而到起点的距离 d_s 为期望值 $\sum_{b=0}^{B} b\,\tilde{P}_s(b)$。

残差与该任务中的其他任务头略有不同：它接收相同的输入，但只共享前两层。残差用于修正距离留下的误差，而不同层级的误差各不相同，因此最后一层针对每个层级分别设置。更具体地说，关于该误差，d_s，$d_e \geq 0$ 会强制预测区间包含 t，而长度短于步幅 q_l 的伪造区间可能在 t 之后开始，或在其之前结束。因此，我们提出了一个残差头来输出 $F_r \in \mathbb{R}^{T_l \times 2}$。残差 $r = 0.5\tanh(F_r[t])$ 可以将起点向右推过 t，或将终点向左推过 t，最多移动半个步幅。

结合距离和残差，公式（1）给出了预测区间 $(\hat{s}_t,\ \hat{e}_t)$ 。然后，我们可以通过比较在 t 处预测的区间与分配给 t的参考伪造区间 g来训练各任务头。如图 1 所示，采用了 Huber 损失 $\mathcal{L}_{\mathrm{hub}}$ 和 IoU 损失 $\mathcal{L}_{\mathrm{IoU}}$：

$$\mathcal{L}_{\mathrm{hub}} = \frac{1}{|\mathcal{P}|}\sum_{t\in\mathcal{P}}\Big[h\big(\tfrac{|\hat{s}_t - s_g|}{\tau_b}\big) + h\big(\tfrac{|\hat{e}_t - e_g|}{\tau_b}\big)\Big], \tag{4}$$

$$\mathcal{L}_{\mathrm{IoU}} = \frac{1}{|\mathcal{P}|}\sum_{t\in\mathcal{P}}(1 - \mathrm{IoU}_t), \tag{5}$$

其中，$h(u) = u^2/2$ 为 $u < 1$，$u - \frac{1}{2}$ 为其他情况时，Huber 函数为 [14]；$\tau_b = 20$ 毫秒是帧分辨率，而 IoU_t 是 $(\hat{s}_t\hat{e}_t)$ 与 g 之间的交并比。两种损失相辅相成：$\mathcal{L}_{\mathrm{IoU}}$ 衡量 $(\hat{s}_t\hat{e}_t)$ 与 g 的重叠程度，但在较长区间上，几毫秒的误差几乎不会改变它。为了使预测更加精确，$\mathcal{L}_{\mathrm{hub}}$ 会根据绝对时间惩罚每个边界处的误差，而不受区间长度影响。

话语级预测。 此任务预测一个话语是否包含任何伪造区间。到目前为止，$\mathcal{L}_{\mathrm{hub}}$、$\mathcal{L}_{\mathrm{IoU}}$ 以及 $\mathcal{L}_{\mathrm{cls}}$ 的第一项都只在正样本时刻上计算。真实话语没有正样本时刻；对于真实数据，反欺骗模型只通过 $\mathcal{L}_{\mathrm{cls}}$ 的第二项进行训练，其中每个时刻的焦点损失只要分数适度偏低，就会变得非常小。

因此，没有任何损失函数会告知反欺骗模型，某个话语整体上是真实语音；对于真实语音输入，反欺骗模型仍会输出候选区间。因此，我们添加了一个话语级头。将六个 SGP 模块的输出沿时间维度拼接，并通过注意力统计池化 [15] 得到加权均值和加权标准差。一个隐藏维度为 256、使用 ReLU 的两层多层感知机，再接一个 Sigmoid函数，将池化后的向量映射为概率 p_u。该任务头使用类别平衡二元交叉熵损失 $\mathcal{L}_{\mathrm{utt}}$进行训练。推理时，若该任务头判定某个话语为真实语音，则不输出任何候选区间。

帧级预测。 这项辅助任务为每一帧预测一个欺骗分数；它用于与帧级对应方法进行比较。该任务头以 SGP 模块第 1 层的输出作为输入，其中一个时刻对应一帧。它与时刻分类头的结构相同。该任务使用二元交叉熵 $\mathcal{L}_{\mathrm{fr}}$ 针对量化帧标签进行训练。梯度在任务头的输入处停止，因此量化标签不会影响 SGP 模块或其他任务头。

最后，总损失为 $\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_h\mathcal{L}_{\mathrm{hub}} + \mathcal{L}_{\mathrm{IoU}} + \mathcal{L}_{\mathrm{utt}} + \mathcal{L}_{\mathrm{fr}}$，其中 $\lambda_h = 0.05$ 在训练起点将 $\mathcal{L}_{\mathrm{hub}}$ 调整到 $\mathcal{L}_{\mathrm{IoU}}$ 的尺度。与其他时序动作检测任务类似， [12, 16,10], 我们采用 Soft-NMS [17]将近似重复的候选区间合并为包含 100 个候选区间的列表。

3. 实验设置

训练配置。 所有反欺骗模型均在 PartialSpoof [2], LlamaPartialSpoof [3] 和 PartialEdit E1 [1]上采用相同的方案进行训练：使用 AdamW，XLS-R 的学习率为 10^{-5}， 10^{-4} 其他模型的权重衰减为 0.05；进行为期 5 个 epoch 的线性预热和为期 30 个 epoch 的余弦衰减；每批次包含 8 个完整话语；采用耐心值为 3 的早停策略；并在一块 48 GB RTX A6000 上对前端使用 fp16自动混合精度。超参数在 PartialSpoof开发集上固定，未针对不同数据集进行调优：$\lambda_h = 0.05$，并在 p_u上采用 0.2 的门控阈值。

评估配置。 每个模型使用三个随机种子进行训练，并报告均值。我们使用四项评估指标来评估反欺骗模型：基于点的等错误率、基于范围的等错误率 [9], 边界误差（起止平均绝对误差）和匹配比例（匹配率）。基于点的等错误率按20毫秒分辨率计算。对于区间级反欺骗模型，每一帧取覆盖该帧的候选区间中的最高分数；若没有候选区间覆盖该帧，则记为0。起止平均绝对误差是预测起止时间与参考值之间的平均绝对误差，单位为毫秒。匹配率是反欺骗模型检测到的参考伪造区间所占的百分比。如果参考区间的最佳预测区间的交并比 ≥ 0.5，则认为该参考区间匹配；平均绝对误差则根据参考区间与交并比最高的预测区间计算。帧级反欺骗模型不输出区间，因此会将归一化伪造分数高于0.5的连续帧拼接成一个预测区间。

4. 结果

比较的反欺骗模型。 针对该任务已发表的多数反欺骗模型都是帧级分类器，因此我们将其与其中几种模型进行比较：BAM（基于 WavLM 特征的边界感知注意力） [7], Split and Conquer（先进行边界检测，再进行片段分类）[6], 以及 MRM（多分辨率反欺骗模型，是唯一报告基于范围的等错误率的反欺骗模型）[2, 9]。与我们的方法最接近的基于候选区间的反欺骗模型是 CFPRF（基于 XLS-R 的由粗到细的提议细化） [5]。

我们的设置。1) 无门控，包含所有候选区间。2) 有门控，话语级头判定为真实语音的话语不会获得候选区间。3) 帧头，仅使用帧级头进行推理。4) 最大融合，对门控提议分数和帧头分数取逐帧最大值。

4.1. 与已发表的 CM 在 PartialSpoof 上的比较

表 1 展示了 PartialSpoof 上的结果。不含话语级头 这一行表示未使用话语级头训练的定位器；它衡量了添加该头的影响。没有该头时，基于点和基于范围的等错误率都会上升，因此联合训练本身有助于定位。推理期间的门控不会改变性能，但它具有

表 1。PartialSpoof 上的结果。†：我们的重新实现。

模型	点 EER	范围等错误率	平均绝对误差（起始/结束）	匹配率
MRM[2, 9]	12.84	30.40	–	–
MRM，重新实现†	15.57	16.07	13.9 / 15.2	69.0
CFPRF[18]	7.61	–	–	–
SaC[6]	6.55	–	–	–
BAM[7]	5.20	–	–	–
Ours, 不含话语	6.01	6.89	9.8 / 9.3	**95.8**
我们的方法，未门控	5.00	5.55	**8.7 / 9.3**	94.5
我们的方法，门控	5.00	5.56	**8.7 / 9.3**	94.3
Ours, 帧头	5.89	6.36	14.2 / 17.9	78.9
Ours, 最大融合	**4.92**	**5.34**	13.0 / 14.3	80.6

表2。LlamaPartialSpoof 上基于点的等错误率

	零样本			微调		
模型	full	部分	both	full	部分	both
MRM[2, 18]	40.98	32.76	44.54	6.04	10.31	8.20
CFPRF[5, 18]	36.22	35.01	41.47	2.89	8.72	5.96
我们的方法，未门控	30.01	26.18	34.39	3.23	1.76	2.85
我们的方法，门控	30.43	26.10	34.63	3.04	**1.71**	2.78
Ours, 帧头	**28.75**	**24.47**	33.84	0.83	2.00	**1.31**
我们的方法，最大融合	29.57	24.60	**33.79**	0.87	1.77	1.33
Ours, 话语头	37.03	–	–	**0.61**	–	–

实际价值：它可以过滤掉真实数据，并减少后处理时间。

每一行中，基于范围的等错误率都高于基于点的等错误率，因为前者还会统计帧内的边界误差。与基线方法相比，我们的反欺骗模型在几乎所有设置下，基于点和基于范围的等错误率都更低。最大融合 效果最好；它优于单独使用候选区间或单独使用帧级输出，说明这两种输出具有互补性。接下来，我们比较起止平均绝对误差和匹配率。回归定位器（无门控、有门控）能将两个边界都定位在 10 ms 以内，并匹配约 95% 的参考区间；相比之下，帧级反欺骗模型（帧头、最大融合）仅能匹配 78–80%，且误差超过 13 ms。我们认为，这种差异源于训练信号：定位器根据精确的参考值进行学习，不会受到量化误差的影响。

4.2. 向 LlamaPartialSpoof 的泛化

LlamaPartialSpoof (LPS) 用作域外测试：评估在 PartialSpoof 上训练的 CM（零样本），或在 LPS 训练子集上微调后的 CM。按照 [18]，相同的设置，我们报告其测试子集三个组的基于点的等错误率：完整，真实语音加完全伪造语音；部分，真实语音加部分伪造语音；以及 两者，所有话语。

表2展示了这些实验中使用的所有 CM 都呈现的两种趋势。首先，零样本评估效果不佳：基于点的等错误率介于 24.47% 和 44.54% 之间，因此仅在 PartialSpoof 上训练的反欺骗模型无法直接应用于 LPS。其次，微调大幅降低了错误率。在两者上，CFPRF 从

从 41.47% 降至 5.96%，而我们的帧头 错误率则从 33.84% 降至 1.31%。在这两种趋势下，我们的反欺骗模型在零样本和微调条件下的错误率都更低。在零样本条件下，我们的设置在每个组别上的结果都更好，优势为 5.79–12.23%，因此，在精确边界上训练的模型泛化能力更好。微调后，我们的方法取得了最佳结果：部分为 1.71–2.00%，两者为 1.31–2.85%。

表 3。在 PartialEdit E1 评估集上的结果，以重新对齐的参考值为准进行评分。†：根据 [1], 中已发布标签和作者自己的评估集报告。

模型	点 EER	范围等错误率	平均绝对误差（起始/结束）	匹配率
基于已发布标签训练				
BAM†[7, 1]	3.80	–	–	–
MRM，重新实现	7.43	7.51	10.7 / 127.3	89.4
我们的方法，未门控	7.23	7.65	25.2 / 82.7	99.0
在重新对齐的标签上训练				
MRM，重新实现	0.53	0.54	10.0 / 8.4	99.0
我们的方法，未门控	1.05	0.78	**7.2 / 5.8**	**99.5**
我们的方法，门控	1.05	0.79	**7.2 / 5.8**	99.4
Ours, 帧头	**0.41**	0.44	8.4 / 6.8	99.1
Ours, 最大融合	0.44	**0.43**	9.6 / 7.4	99.3

有一项例外：无门控 在完整场景下的表现不如 CFPRF。进一步调查后，我们发现这是因为候选区间并不总能覆盖整个伪造话语，这表明任务存在不匹配（话语级检测与部分伪造定位）。单独使用话语级头（第话语头行）可达到 0.61%，这表明学习到的特征具有鲁棒性。

4.3. 揭示标签错位问题

表 3 展示了 PartialEdit E1 上的结果。使用已发布标签进行训练后，我们的反欺骗模型的基于点的等错误率高于 BAM†。作为第二个参考，我们纳入了 MRM 的一个重新实现版本。两个反欺骗模型有一个共同点：终点误差是起点误差的数倍。为检查标签，我们通过归一化互相关（NCC），比较每个包含单个伪造区间的文件与其 VCTK 源文件 [19]，如图 3 所示。就起点而言，已发布的时间标签相当准确，92% 的文件中，编辑后的文件与源文件相匹配（NCC $>$ 0.99）。然而，就终点标签而言，匹配文件的比例仅为 37%。

因此，我们重新对齐了所有单区间文件，并从 PartialEdit E1[3] 中创建了一个“重新对齐”子集。使用该子集进行训练后，两种反欺骗模型的基于点的等错误率均从 7% 以上降至约 1% 或更低，而且我们的反欺骗模型在 起止平均绝对误差 和匹配率方面表现最佳。参考模型 MRM 也受益于使用重新对齐的训练数据。

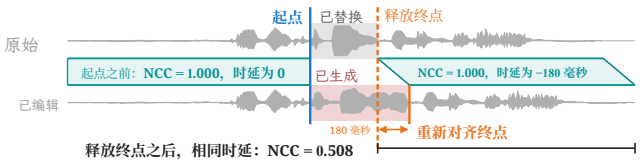

图 3。使用互相关比较边界。

5. 结论

我们提出了一种单阶段反欺骗模型，将欺骗定位视为边界回归，根据精确的参考值预测起始时间和结束时间，无需量化。该模型在 PartialSpoof 上实现了最低的基于点和基于范围的等错误率，并且对 LlamaPartialSpoof 的泛化能力更强。边界误差分析促使我们对 PartialEdit 进行了重新对齐。

[3]https://huggingface.co/datasets/DeepFense/partialedit-corrected

6. 遵守伦理标准

本研究使用了开放获取的语音语料库（PartialSpoof、LlamaPartialSpoof、PartialEdit 和 VCTK）。根据开放获取数据所附的许可协议，本研究无需获得伦理审批。

7. 参考文献

[1] You Zhang、Baotong Tian、Lin Zhang 和 Zhiyao Duan，“PartialEdit：神经语音编辑时代的部分深度伪造识别”，载于*Interspeech* 论文集，2025 年，第 5353–5357 页。[2] Lin Zhang、Xin Wang、Erica Cooper、Nicholas Evans 和 Junichi Yamagishi，“PartialSpoof 数据库及用于检测话语中嵌入的短伪造语音片段的反欺骗措施”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 31 卷，第 813–825 页，2023 年。[3] Hieu-Thi Luong、Haoyang Li、Lin Zhang、Kong Aik Lee 和 Eng Siong Chng，“LlamaPartialSpoof：模拟虚假信息生成的大语言模型驱动的伪造语音数据集”，载于*ICASSP* 论文集，2025 年，第 1–5 页。[4] Arun Babu、Changhan Wang、Andros Tjandra、Kushal Lakhotia、Qiantong Xu、Naman Goyal、Kritika Singh、Patrick von Platen、Yatharth Saraf、Juan Pino、Alexei Baevski、Alexis Conneau 和 Michael Auli，“XLS-R：大规模自监督跨语言语音表征学习”，载于*Interspeech* 论文集，2022 年，第 2278–2282 页。[5] Junyan Wu、Wei Lu、Xiangyang Luo、Rui Yang、Qian Wang 和 Xiaochun Cao，“用于音频时序伪造检测与定位的由粗到细的候选细化框架”，载于 *ACM* 多媒体国际会议论文集（*ACM MM*），2024 年，第 7395–7403 页。[6]Inbal Rimon、Oren Gal 和 Haim Permuter，“ Split and Conquer 部分语音深度伪造”，*arXiv* 预印本 *arXiv:2604.02913*，2026 年。[7] Jiafeng Zhong、Bin Li 和 Jiangyan Yi，“利用边界感知注意力机制增强部分欺骗音频的定位”，载于*Interspeech* 论文集，2024 年，第 4838–4842 页。[8] Yuchen Mao、Wen Huang 和 Yanmin Qian，“通过片段感知学习实现超越过渡区域的语音深度伪造定位”，载于*ICASSP* 论文集，2026 年，第 16507–16511 页。[9] Lin Zhang、Xin Wang、Erica Cooper、Nicholas Evans 和 Junichi Yamagishi，“用于欺骗定位的基于范围的等错误率”，载于*Interspeech* 论文集，2023 年，第 3212–3216 页。[10] Dingfeng Shi、Yujie Zhong、Qiong Cao、Lin Ma、Jia Li 和 Dacheng Tao，“TriDet：采用相对边界建模的时序动作检测”，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集（*CVPR*），2023 年，第 18857–18866 页。[11] Zhi Tian、Chunhua Shen、Hao Chen 和 Tong He，“FCOS：全卷积单阶段目标检测”，载于*IEEE/CVF* 计算机视觉国际会议论文集（*ICCV*），2019 年，第 9626–9635 页。[12] Chen-Lin Zhang、Jianxin Wu 和 Yin Li，“ActionFormer：利用变换器定位动作时刻”，载于欧洲计算机视觉会议论文集（*ECCV*），2022 年，第 492–510 页。[13] Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaiming He 和 Piotr Doll´ar，“用于密集目标检测的焦点损失”，载于*IEEE* 计算机视觉国际会议论文集（*ICCV*），2017 年，第 2999–3007 页。

[14] Gregory P Meyer，《Huber 损失的另一种概率解释》，载于*CVPR*论文集。IEEE，2021年，第5257–5265页。[15] Koji Okabe、Takafumi Koshinaka和Koichi Shinoda，《用于深度说话人嵌入的注意力统计池化》，载于*Interspeech*论文集，2018年，第2252–2256页。[16] Tianwei Lin、Xiao Liu、Xin Li、Errui Ding和Shilei Wen，《BMN：用于生成时序动作提议的边界匹配网络》，载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，2019年，第3888–3897页。[17] Navaneeth Bodla、Bharat Singh、Rama Chellappa和Larry S. Davis，《 Soft-NMS：仅用一行代码改进目标检测》，载于*IEEE*国际计算机视觉会议（*ICCV*）论文集，2017年，第5562–5570页。[18] Hieu-Thi Luong、Inbal Rimon、Haim Permuter、Kong Aik Lee和Eng Siong Chng，《部分伪造语音的鲁棒定位：评估指标与域外评估》，载于*APSIPA*年度峰会与会议（*APSIPA ASC*）论文集，2025年，第2205–2210页。[19] Junichi Yamagishi、Christophe Veaux和Kirsten MacDonald，《CSTR VCTK语料库：用于CSTR语音克隆工具包的英语多说话人语料库（版本0.92）》，爱丁堡大学语音技术研究中心，2019年。[20] Niko Br¨ummer、Luciana Ferrer和Albert Swart，《一百次试验中，你的说话人验证器会产生多少次错误？》，载于*Interspeech*论文集，2021年，第1059–1063页。[21] Richard O. Duda、Peter E. Hart和David G. Stork，《模式分类》，Wiley，纽约，第2版，2001年。[22] NIST，《NIST 2016说话人识别评估计划》，2016年。

A. 解读部分伪造定位的评估指标

本附录回顾传统二分类中等错误率（EER）的计算，并将其扩展至部分伪造检测与定位的评估指标。内容如下：

1. 关于二元分类任务（§ A.1）：正如 [20], 所指出的，在分数经过完美校准的情况下，等错误率是贝叶斯决策代价关于先验概率的最大值。

2. 关于部分伪造（§A.2）：

- 分段等错误率（§A.2.1）：采用与传统二元分类中的等错误率相同的方式进行解读，前提是分段分数独立同分布。
- 范围等错误率（§A.2.2）：假设分数 $r(t)$ 是时间的函数。以类似于传统等错误率的方式解释范围等错误率，需要对分数分布作出强假设。

A.1. 二元分类

A.1.1. 最优决策

二元分类器的用户需要判断输入波形 $\boldsymbol{x} \in \mathcal{X}$ 是真实还是伪造。该决策为二元决策，目标标签可以是 $\mathcal{Y} = \{H_{\mathsf{Y}}, H_{\mathsf{N}}\}$，其中

$$H_{\mathsf{Y}} : \boldsymbol{x} \text{ is REAL},$$
$$H_{\mathsf{N}} : \boldsymbol{x} \text{ is FAKE}.$$

用户操作 a 可以是 $\mathcal{A} = \{$接受，拒绝 $\mathtt{T}\}$:

$$\mathtt{ACCEPT} : \text{user claims that } \boldsymbol{x} \text{ is REAL},$$
$$\mathtt{REJECT} : \text{user claims that } \boldsymbol{x} \text{ is FAKE}.$$

用户在作出决策时会使用检测系统；$\boldsymbol{x}$ 会被处理并转换为分数 $s \in \mathbb{R}$，其数值表示该 $\boldsymbol{x}$ 是真实的可能性有多大。直观来说，如果 s足够高，表明其为真实的可能性较大，用户就应该采取$a =$ 接受操作。

用户可能采取错误的行动：真实样本（H_{Y}）可能被错误拒绝（漏检），而伪造样本（H_{N}）可能被错误接受（误接受）。错误行动可以分配决策代价。给定行动 $a \in \mathcal{A}$ 和真实标签$y \in \mathcal{Y}$，我们使用 $C(a, y) \geq 0$ 表示决策成本函数 c： $\mathcal{A} \times \mathcal{Y} \to \mathbb{R}^+$。表4给出了一个示例。

根据贝叶斯决策 [21], 的实践，当给定输入 s 采取行动 a 时，用户的条件风险 $R(a|s)$ 计算如下：

$$R(a|s) = \sum_{y \in \mathcal{Y}} C(a, y) P(y|s), \tag{6}$$

其中， $P(y|s)$ 是观测到 s 时类别 y 的后验概率。将表4中列出的 $C(a, y)$ 的可能取值代入后，我们可以计算采取两种可能行动时的条件风险

$$R(\mathtt{ACCEPT}|s) = C_{\text{fa}}^{\mathsf{N}} P(H_{\mathsf{N}}|s), \tag{7}$$
$$R(\mathtt{REJECT}|s) = C_{\text{fr}}^{\mathsf{Y}} P(H_{\mathsf{Y}}|s). \tag{8}$$

表4。决策代价 $C(a, y)$ 的取值。下标 fa 和 fr 分别表示错误接受和错误拒绝。

	行动 =接受	行动 = 拒绝
真实类别 =Y	0	$C_{\text{fr}}^{\mathsf{Y}}$
真实类别 =N	$C_{\text{fa}}^{\mathsf{N}}$	0

贝叶斯决策规则要求用户选择风险较小的行动 a [21],，即，

$$a = \begin{cases} \mathtt{ACCEPT}, & \text{if } R(\mathtt{REJECT}|s) > R(\mathtt{ACCEPT}|s) \\ \mathtt{REJECT}, & \text{if } R(\mathtt{REJECT}|s) \leq R(\mathtt{ACCEPT}|s) \end{cases}. \tag{9}$$

利用式(7)和式(8)，上述策略可改写为

$$C_{\text{fr}}^{\mathsf{Y}} P(H_{\mathsf{Y}}|s) \underset{\mathtt{REJECT}}{\overset{\mathtt{ACCEPT}}{\gtrless}} C_{\text{fa}}^{\mathsf{N}} P(H_{\mathsf{N}}|s). \tag{10}$$

根据贝叶斯公式，用户知道 $P(H_{\mathsf{Y}}|s) = p(s|H_{\mathsf{Y}})\pi_{\mathsf{Y}}/p(s)$ 且 $P(H_{\mathsf{N}}|s) = p(s|H_{\mathsf{N}})\pi_{\mathsf{N}}/p(s)$，其中 $\pi_{\mathsf{Y}} \triangleq P(H_{\mathsf{Y}}) = 1 - \pi_{\mathsf{N}}$ 是 H_{Y} 的先验概率。随后，用户可以进一步将该策略改写为对数比形式：

$$\log \frac{p(s|H_{\mathsf{Y}})}{p(s|H_{\mathsf{N}})} \underset{\mathtt{REJECT}}{\overset{\mathtt{ACCEPT}}{\gtrless}} \log\Big(\frac{(1-\pi_{\mathsf{Y}})}{\pi_{\mathsf{Y}}} \frac{C_{\text{fa}}^{\mathsf{N}}}{C_{\text{fr}}^{\mathsf{Y}}}\Big). \tag{11}$$

式(11)左侧的项是观测到的 s 的对数似然比（LLR），LLR(s) $\triangleq \log \frac{p_{(s|H_{\mathsf{Y}})}}{p(s|H_{\mathsf{N}})}$。这是唯一一个依赖输入语音数据的项（通过 s）。logit $\log \frac{\pi_{\mathsf{Y}}}{(1-\pi_{\mathsf{Y}})}$ 以及决策代价项由用户预先设定。换言之，用户启动检测系统时，会配置先验概率和决策代价。给定输入 $\boldsymbol{x}$，检测器会生成 s，用户据此按照式(11)作出决策。

这是用户的最优决策策略。

A.1.2. 等错误率的解读

当用户拥有理想的对数似然比及其分布时。令 $r \triangleq$LLR(s) $\triangleq \log \frac{p_{(s|H_{\mathsf{Y}})}}{p(s|H_{\mathsf{N}})}$ 作为对应于 s 的对数似然比值的简写。代入决策代价后，当用户遵循式 (11) 中的最优贝叶斯决策策略时，总体风险 $R(\pi_{\mathsf{Y}}, C_{\text{fa}}^{\mathsf{N}}, C_{\text{fr}}^{\mathsf{Y}})$ 变为：

$$R(\pi_{\mathsf{Y}}, C_{\text{fa}}^{\mathsf{N}}, C_{\text{fr}}^{\mathsf{Y}}) = \sum_{a \in \mathcal{A}} \int R(a|r) p(r) dr \tag{12}$$

$$= \int \sum_{a \in \mathcal{A}} \sum_{y \in \mathcal{Y}} C(a, y) P(y|r) p(r) dr \tag{13}$$

$$= \int_{\tau_{\text{Bayes}}}^{+\infty} C_{\text{fa}}^{\mathsf{N}} P(H_{\mathsf{N}}|r) p(r) dr + \int_{-\infty}^{\tau_{\text{Bayes}}} C_{\text{fr}}^{\mathsf{Y}} P(H_{\mathsf{Y}}|r) p(r) dr \tag{14}$$

$$= C_{\text{fa}}^{\mathsf{N}}(1 - \pi_{\mathsf{Y}}) \int_{\tau_{\text{Bayes}}}^{+\infty} p(r|H_{\mathsf{N}}) dr + C_{\text{fr}}^{\mathsf{Y}} \pi_{\mathsf{Y}} \int_{-\infty}^{\tau_{\text{Bayes}}} p(r|H_{\mathsf{Y}}) dr \tag{15}$$

$$= C_{\text{fa}}^{\mathsf{N}}(1 - \pi_{\mathsf{Y}}) p_{\text{fa}}(\tau_{\text{Bayes}}) + C_{\text{fr}}^{\mathsf{Y}} \pi_{\mathsf{Y}} p_{\text{miss}}(\tau_{\text{Bayes}}), \tag{16}$$

其中，$p_{\text{fa}}(\tau_{\text{Bayes}})$ 和 $p_{\text{miss}}(\tau_{\text{Bayes}})$ 分别对应错误接受和错误拒绝的概率，

$$p_{\text{fa}}(\tau_{\text{Bayes}}) \triangleq \int_{\tau_{\text{Bayes}}}^{+\infty} p(r|H_{\text{N}})dr, \tag{17}$$

$$p_{\text{miss}}(\tau_{\text{Bayes}}) \triangleq \int_{-\infty}^{\tau_{\text{Bayes}}} p(r|H_{\text{Y}})dr, \tag{18}$$

其中，$\tau_{\text{Bayes}} = \log\Big((1-\pi_{\text{Y}})C_{\text{fa}}^{\text{N}}/(\pi_{\text{Y}}C_{\text{fr}}^{\text{Y}})\Big)$ 是由用户声明的先验概率和决策代价确定的最优贝叶斯决策阈值。[4]

总体风险 R 是 $\{\pi_{\text{Y}}, C_{\text{fa}}^{\text{N}}, C_{\text{fr}}^{\text{Y}}\}$的函数这一概念反映了这样一个事实：给定 r 及其分布后，用户的决策风险仅受这三个由用户声明的参数影响。如果用户更改这些参数，决策阈值 τ_{Bayes} 以及相应的总体风险也会自动改变。

已证明，[20, § 2], 在给定 $C_{\text{fa}}^{\text{N}} = C_{\text{fr}}^{\text{Y}} = 1$的情况下，当用户改变 π_{Y}时，等错误率是最大的总体风险。

$$\text{EER} = \max_{\pi_{\text{Y}}} R(\pi_{\text{Y}}, C_{\text{fa}}^{\text{N}} = 1, C_{\text{fr}}^{\text{Y}} = 1) \tag{19}$$

$$= \max_{\pi_{\text{Y}}} \Big((1-\pi_{\text{Y}})p_{\text{fa}}(\tau(\pi_{\text{Y}})) + \pi_{\text{Y}}p_{\text{miss}}(\tau(\pi_{\text{Y}}))\Big), \tag{20}$$

其中，阈值 $\tau(\pi_{\text{Y}}) = \log\frac{1-\pi_{\text{Y}}}{\pi_{\text{Y}}}$是 π_{Y}的函数。使总体风险达到最大值的π_{Y}^* 对应于阈值 $\tau_{\text{EER}} = \log\frac{1-\pi^*_{\text{Y}}}{\pi^*_{\text{Y}}}$，该阈值还满足 $p_{\text{fa}}(\tau_{\text{EER}}) = p_{\text{miss}}(\tau_{\text{EER}})$，因此得名等错误率。

这里，我们再次写出证明。令 $\text{R}(\pi_{\text{Y}}, C_{\text{fa}}^{\text{N}}, C_{\text{fr}}^{\text{Y}}) \triangleq R(\pi_{\text{Y}}, C_{\text{fa}}^{\text{N}} = 1, C_{\text{fr}}^{\text{Y}} = 1)$ 表示决策风险的简写。对 π_{Y} 求一阶导数，可得

$$\frac{d\text{R}(\pi_{\text{Y}})}{d\pi_{\text{Y}}} = -p_{\text{fa}}(\tau(\pi_{\text{Y}})) + (1-\pi_{\text{Y}})\frac{dp_{\text{fa}}(\tau(\pi_{\text{Y}}))}{d\pi_{\text{Y}}} \tag{21}$$

$$+ p_{\text{miss}}(\tau(\pi_{\text{Y}})) + \pi_{\text{Y}}\frac{dp_{\text{miss}}(\tau(\pi_{\text{Y}}))}{d\pi_{\text{Y}}}, \tag{22}$$

其中两个错误率的导数为

$$\frac{dp_{\text{fa}}(\tau(\pi_{\text{Y}}))}{d\pi_{\text{Y}}} = -p(\tau(\pi_{\text{Y}})|H_{\text{N}})\frac{d\tau(\pi_{\text{Y}})}{d\pi_{\text{Y}}} \tag{23}$$

$$\frac{dp_{\text{miss}}(\tau(\pi_{\text{Y}}))}{d\pi_{\text{Y}}} = p(\tau(\pi_{\text{Y}})|H_{\text{Y}})\frac{d\tau(\pi_{\text{Y}})}{d\pi_{\text{Y}}}, \tag{24}$$

并且

$$\frac{d\tau(\pi_{\text{Y}})}{d\pi_{\text{Y}}} = \frac{-1}{(1-\pi_{\text{Y}})\pi_{\text{Y}}}. \tag{25}$$

由于 LLR 的 LLR 仍是 LLR [20], 即，$r = \log\frac{p(r|H_{\text{Y}})}{p(r|H_{\text{N}})}$，因此在任意决策阈值 $r = \tau(\pi_{\text{Y}})$ 下，我们知道

$$p(\tau(\pi_{\text{Y}})|H_{\text{Y}}) = \exp(\tau(\pi_{\text{Y}}))p(\tau(\pi_{\text{Y}})|H_{\text{N}}) = \frac{1-\pi_{\text{Y}}}{\pi_{\text{Y}}}p(\tau(\pi_{\text{Y}})|H_{\text{N}}). \tag{26}$$

[4]第（15）行成立是因为 $P(H_{\text{N}}|r)p(r) = (1-\pi_{\text{Y}})p(r|H_{\text{N}})$ 且 $P(H_{\text{Y}}|r)p(r) = \pi_{\text{Y}}p(r|H_{\text{Y}})$。此外，先验概率和决策代价独立于 r，因此可移至积分号外。

利用上述四个方程，可以证明

$$(1-\pi_{\text{Y}})\frac{dp_{\text{fa}}(\tau(\pi_{\text{Y}}))}{d\pi_{\text{Y}}} + \pi_{\text{Y}}\frac{dp_{\text{miss}}(\tau(\pi_{\text{Y}}))}{d\pi_{\text{Y}}} \tag{27}$$

$$= -(1-\pi_{\text{Y}})p(\tau(\pi_{\text{Y}})|H_{\text{N}})\frac{d\tau(\pi_{\text{Y}})}{d\pi_{\text{Y}}} + \pi_{\text{Y}}p(\tau(\pi_{\text{Y}})|H_{\text{Y}})\frac{d\tau(\pi_{\text{Y}})}{d\pi_{\text{Y}}} \tag{28}$$

$$= \Big(-(1-\pi_{\text{Y}}) + \pi_{\text{Y}}\frac{(1-\pi_{\text{Y}})}{\pi_{\text{Y}}}\Big)p(\tau(\pi_{\text{Y}})|H_{\text{N}})\frac{d\tau(\pi_{\text{Y}})}{d\pi_{\text{Y}}} \tag{29}$$

$$= 0 \tag{30}$$

因此，式（22）中的第2项和第4项相互抵消，式（22）变为

$$\frac{d\text{R}(\pi_{\text{Y}})}{d\pi_{\text{Y}}} = p_{\text{miss}}(\tau(\pi_{\text{Y}})) - p_{\text{fa}}(\tau(\pi_{\text{Y}})). \tag{31}$$

求二阶导数，可得

$$\frac{d^2\text{R}(\pi_{\text{Y}})}{d\pi_{\text{Y}}^2} = \Big(p(\tau(\pi_{\text{Y}})|H_{\text{Y}}) + p(\tau(\pi_{\text{Y}})|H_{\text{N}})\Big)\frac{-1}{(1-\pi_{\text{Y}})\pi_{\text{Y}}} < 0. \tag{32}$$

因此，当变量 π_{Y}^* 满足 $\frac{d\text{R}(\pi_{\text{Y}})}{d\pi_{\text{Y}}}\big|_{\pi_{\text{Y}}=\pi_{\text{Y}}*} = p_{\text{miss}}(\tau(\pi_{\text{Y}}^*)) - p_{\text{fa}}(\tau(\pi_{\text{Y}}^*)) = 0$ 时，$\text{R}(\pi_{\text{Y}}^*)$ 取得全局最大值。换言之，决策阈值 $\tau(\pi_{\text{Y}}^*) = \log\frac{1-\pi^*_{\text{Y}}}{\pi^*_{\text{Y}}}$ 使两种错误率相等。——

简言之，等错误率是相对于用户声明先验概率 π_{Y} 的最大风险。上述解释要求

1. 决策基于完美的对数似然比 r，而非未经校准的分数。.
2. 决策基于贝叶斯决策策略 (11) .
3. 风险 R 假设两类决策错误的决策代价固定且相等。

用户拥有理想的对数似然比，但不了解其分布时。

实际上，用户并不知道真实的 $p(r|H_{\text{Y}})$ 和 $p(r|H_{\text{N}})$，因此无法通过积分计算错误接受和拒绝的概率。用户必须通过计数，在测试集上估计错误概率。

设 $\mathcal{D}_{r,\text{Y}} = \{r_i|H_{\text{Y}}\}$ 和 $\mathcal{D}_{r,\text{N}} = \{r_i|H_{\text{N}}\}$ 分别为测试集中 H_{Y} 样本和 H_{N} 样本的对数似然比。设一般阈值为 τ，它可能不同于最优阈值 τ_{Bayes}。错误率估计如下：

$$\widehat{P}_{\text{fa}}(\tau) = \frac{1}{|\mathcal{D}_{r,\text{N}}|}\sum_{r_i\in\mathcal{D}_{r,\text{N}}} 1(r_i > \tau), \tag{33}$$

$$\widehat{P}_{\text{miss}}(\tau) = \frac{1}{|\mathcal{D}_{r,\text{Y}}|}\sum_{r_i\in\mathcal{D}_{r,\text{Y}}} 1(r_i < \tau), \tag{34}$$

其中，$1(\cdot)$是指示函数。将估计的错误率代入后，用户得到估计的总体风险

$$\widehat{R}(\pi_{\text{Y}}, C_{\text{fa}}^{\text{N}}, C_{\text{fr}}^{\text{Y}}, \tau) = C_{\text{fa}}^{\text{N}}(1-\pi_{\text{Y}})\widehat{P}_{\text{fa}}(\tau) + C_{\text{fr}}^{\text{Y}}\pi_{\text{Y}}\widehat{P}_{\text{miss}}(\tau). \tag{35}$$

请注意，这就是决策代价函数（DCF），例如 NIST SRE [22, Eq.(2)]中使用的函数。如果阈值按照贝叶斯决策规则 $\tau_{\text{Bayes}} = \log\Big((1-\pi_{\text{Y}})C_{\text{fa}}^{\text{N}}/(\pi_{\text{Y}}C_{\text{fr}}^{\text{Y}})\Big)$设定，那么所得的 DCF 是对式（16）中贝叶斯风险的近似，其中两类错误率是通过计数而非积分近似得到的。

因此，当 $C_{\text{fa}}^{\text{N}} = C_{\text{fr}}^{\text{Y}} = 1$时，等错误率的性质成立。然而，由于 $P_{\text{fa}}(\tau)$ 和 $P_{\text{miss}}(\tau)$ 不再是 τ 的连续函数

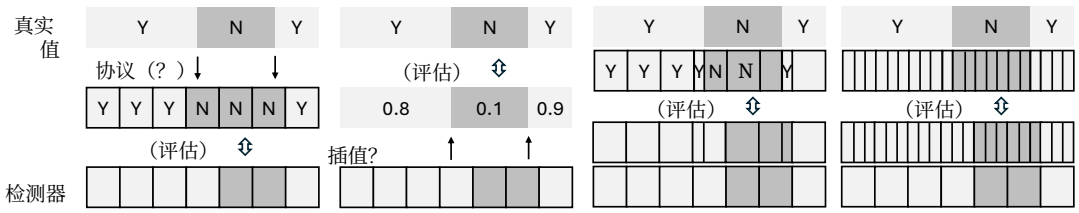

图4。部分欺骗检测评估中，解决真实值与模型输出之间对齐问题的三种方法。

，可能无法找到一个 τ_{EER}，使得 $P_{\text{fa}}(\tau_{\text{EER}})$ 恰好等于 $P_{\text{miss}}(\tau_{\text{EER}})$。作为近似，可以将阈值设为

$$\hat{\tau}_{\text{EER}} \approx \arg\min_{\tau} |\widehat{P}_{\text{miss}}(\tau) - \widehat{P}_{\text{fa}}(\tau)|, \tag{36}$$

且等错误率可近似为

$$\widehat{\text{EER}} \approx \left(\widehat{P}_{\text{miss}}(\hat{\tau}_{\text{EER}}) + \widehat{P}_{\text{fa}}(\hat{\tau}_{\text{EER}})\right)/2. \tag{37}$$

上一小节中的相同要求均已满足（或假定为真）。在本文其余部分中，我们还假设检测器能够生成完美的对数似然比。

A.2. 部分欺骗检测

A.2.1. 真实标签与检测器输出之间的对齐

在部分伪造场景中，一段话语由至少一个真实片段和一个伪造片段拼接而成。首要任务是对每个片段作出判断。

这种设置不同于上一节中的话语级二元分类。一个主要原因是，片段的长度是任意的，而检测器可能只能以统一的时间区间生成分数。这会导致评估中的对齐问题。

在讨论处理对齐问题的不同方法之前，我们需要就评估原则达成一致：对于需要真实值和模型输出的评估，无论模型的行为如何，真实值都应保持不变。

处理这一问题可能有三种方式（如图 4 所示）：

1. 转换真实标签，使其能够在相同的均匀时间间隔上分配真实标签。2. 将检测器的输出分数转换为真实标签所定义的每个片段对应的分数。3. 同时转换真实标签和模型输出，使得每当触及真实值的边界或均匀片段的边界时，就插入一个边界。4. 通过以区间 $\Delta > 0$ 对两者进行拆分，同时转换真实标签和模型输出。

我们可以认为，第一种策略并非最优，因为它违背了我们所认同的原则。违背该原则的一个后果是：当固定长度区间内存在歧义时（例如图4左中部最左和最右的“N”），我们需要额外就如何转换真实标签达成“共识”。 [5]

[5]如果评估集附带预先定义的真实标签集合 [2]，则可以在一定程度上达成一致。

此外，真实标签的转换还会受到检测器能力的影响（而这本不应如此）。如果两个检测器分别以每8毫秒和每15毫秒预测一个分数，那么它们的“真实标签”从本质上就不同。使用不同真实值进行的评估结果不可比较。

第 3 种策略同样违背了这一原则。转换后的真实标签取决于检测器的行为。

第 2 种策略更好，因为真实标签始终不会改变。发生变化的是模型的输出——如果检测器每 8 毫秒生成一个分数，就必须包含一个后处理模块，以便为真实标签所定义的每个片段获取一个分数。检测器有责任按照评估指标所要求的格式生成分数。

第 4 种策略使用一个与检测器设计无关的区间 Δ。如果 Δ 足够小，我们就能获得对齐的真实标签和模型输出。

A.2.2. 问题设定

无论选择哪种方式，我们都可以将任务定义概括如下：用户需要猜测或判断输入话语从时间 t_s t_e $\boldsymbol{x} \in \mathcal{X}$ 到某一时刻的输入片段是真实还是伪造的。该判断为二元判断，目标标签可以是$\mathcal{Y} = \{H_{\text{Y}}, H_{\text{N}}\}$，其中

$$H_{\text{Y}} : \boldsymbol{x}(t_s, t_e) \text{ is REAL},$$
$$H_{\text{N}} : \boldsymbol{x}(t_s, t_e) \text{ is FAKE}.$$

用户操作 a 可以是 $\mathcal{A} = \{$接受，拒绝$\}$:

$$\texttt{ACCEPT} : \text{user claims that } \boldsymbol{x}(t_s, t_e) \text{ is REAL},$$
$$\texttt{REJECT} : \text{user claims that } \boldsymbol{x}(t_s, t_e) \text{ is FAKE}.$$

A.2.3. 片段等错误率

最初，片段等错误率 [9] 假设采用第一种对齐策略。因此，每个样本的（t_s， t_e）由检测器的“时间分辨率”定义：即检测器输出分数的时间区间。 $t_e - t_s$ 的值等于“时间分辨率”。贝叶斯决策、决策风险和等错误率可以按照 A.1 节中的话语级分类以相同方式推导。时间信息（t_s， t_e）不会影响错误代价，并且假设所有样本彼此独立。

A.2.4. 范围等错误率

错误拒绝和误接受率。 改用范围等错误率的理由是，在转换真实标签 [9] 时会产生歧义。缓解策略是采用范围等错误率。以下是对这一思路的详细解释。

Range-EER 假设采用第 4 种对齐策略。问题设置与片段等错误率类似，但需要评分的数据变为 $\boldsymbol{x}(t, t+\Delta)$——我们需要一个检测器，能够为 $\boldsymbol{x}(t, t+\Delta)$ $\forall t, \forall \Delta$[6] 生成分数。

不失一般性，假设一段话语中只有一个伪造片段（t_s，t_e）。子片段的数量为 $(t_e - t_s)/\Delta$。[7] 设分数 $r(t)$ 是时间 t 的函数。设决策阈值为 τ，第 i 个子片段的对数似然比（LLR）分数为 $r(t_s + i\Delta)$；于是可以将误接受率计算为

$$\widehat{P}_{\text{fa}}(\tau) = \lim_{\Delta \to 0} \frac{1}{(t_e - t_s)/\Delta} \sum_{i=1}^{\frac{t_e - t_s}{\Delta}} 1(r(t_s + i\Delta) > \tau) \tag{38}$$

$$= \lim_{\Delta \to 0} \frac{1}{(t_e - t_s)} \sum_{i=1}^{\frac{t_e - t_s}{\Delta}} 1(r(t_s + i\Delta) > \tau)\Delta \tag{39}$$

$$= \frac{1}{t_e - t_s} \int_{t_s}^{t_e} 1(r(t) > \tau)\, dt \tag{40}$$

其中，$1(\cdot)$ 是指示函数。当 $\Delta \to$趋于 0 时，根据黎曼和的定义[8]，通过积分可以得到该结果。直观地说，$\widehat{P}_{\text{fa}}(\tau)$ 对应于这样一个区间的比例：在该区间内，分数 $r(t)$ 高于阈值 τ。

直观地说，给定一个评估集，我们可以将所有（真实标签的）伪造片段及其分数函数连接成单个 $r(t)$。误接受的定义很简单：

$$\widehat{P}_{\text{fa}}(\tau) = \frac{1}{T_{\text{N}}} \int_0^{T_{\text{N}}} 1(r(t) > \tau)\, dt, \tag{41}$$

其中，T_{N} 是拼接后伪造片段的长度。类似地，我们可以将错误拒绝率定义为

$$\widehat{P}_{\text{miss}}(\tau) = \frac{1}{T_{\text{Y}}} \int_0^{T_{\text{Y}}} 1(r(t) < \tau)\, dt \tag{42}$$

以上定义与原论文中的定义略有不同，主要有两点：1）将伪造（或真实）片段拼接成一个整体，以简化符号表示，且不会影响结果[9]；2）该定义对 $\Delta \to 0$ 取极限，并将求和转换为积分。当然，实际中我们必须通过对较小的 Δ求和来实现。

贝叶斯决策? 范围等错误率只需找到阈值 τ_{EER}，使得$\widehat{P}_{\text{miss}}(\tau_{\text{EER}}) = \widehat{P}_{\text{fa}}(\tau_{\text{EER}})$。这个定义无可非议。然而，问题在于，在进行贝叶斯决策时，这个操作点 τ_{EER} 是否对应于最大决策风险（相对于先验概率）。

假设未校准分数 $s(t)$，并且针对 t 独立地进行决策，我们可以沿用 § A.1.1 中的最优贝叶斯决策，将策略写为

$$\log \frac{p(s(t)|H_{\text{Y}})}{p(s(t)|H_{\text{N}})} \underset{\text{REJECT}}{\overset{\text{ACCEPT}}{\gtrless}} \log\Big(\frac{(1-\pi_{\text{Y}})}{\pi_{\text{Y}}} \frac{C_{\text{fa}}^{\text{N}}}{C_{\text{fr}}^{\text{Y}}}\Big), \tag{43}$$

其中 $\tau_{\text{Bayes}} \triangleq \log\left(\frac{(1-\pi_{\text{Y}})\, C_{\text{Nfa}} C_{\text{Yfr}}}{\pi}\right)$ 是由先验概率和代价定义的决策阈值。然后我们可以令 $r(t) \triangleq \text{LLR}(s(t)) \triangleq \log \frac{p_{(s(t)|H_{\text{Y}})}}{p(s(t)|H_{\text{N}})}$，并将决策风险（暂时）写成

$$R(\pi_{\text{Y}}, C_{\text{fa}}^{\text{N}}, C_{\text{fr}}^{\text{Y}}) \tag{44}$$

$$= C_{\text{fa}}^{\text{N}}(1-\pi_{\text{Y}}) \frac{1}{T_{\text{N}}} \int_0^{T_{\text{N}}} \int_{\tau_{\text{Bayes}}}^{+\infty} p(r(t)|H_{\text{N}}) dr dt \tag{45}$$

$$+ C_{\text{fr}}^{\text{Y}} \pi_{\text{Y}} \frac{1}{T_{\text{Y}}} \int_0^{T_{\text{Y}}} \int_{-\infty}^{\tau_{\text{Bayes}}} p(r(t)|H_{\text{Y}}) dr dt \tag{46}$$

与传统二元分类的情形（式 (16)）不同，分数随时间变化，因此 $r(t)$ 可视为随机过程。[10]

为简化这个二重积分，我们假设 $r(t)$ 是平稳的：$p(r(t)|H) \equiv p(r|H)$。因此，对 t 的积分会被平均掉：

$$R(\pi_{\text{Y}}, C_{\text{fa}}^{\text{N}}, C_{\text{fr}}^{\text{Y}}) \tag{47}$$

$$= C_{\text{fa}}^{\text{N}}(1-\pi_{\text{Y}}) \int_{\tau_{\text{Bayes}}}^{+\infty} p(r|H_{\text{N}}) dr + C_{\text{fr}}^{\text{Y}} \pi_{\text{Y}} \int_{-\infty}^{\tau_{\text{Bayes}}} p(r|H_{\text{Y}}) dr.$$

在平稳假设下，风险的形式与传统二分类的情况相同（式（16））。与上一节中的证明类似，将导数设为零，即可得到贝叶斯阈值$\tau_{\text{EER}} = \tau(\pi^*_{\text{Y}})$处的最大决策风险，其中：

$$\int_{\tau_{\text{EER}}}^{+\infty} p(r|H_{\text{N}}) dr = \int_{-\infty}^{\tau_{\text{EER}}} p(r|H_{\text{Y}}) dr. \tag{48}$$

接下来，我们将上述积分与 $\widehat{P}_{\text{fa}}(\tau)$和 $\widehat{P}_{\text{miss}}(\tau)$的定义联系起来。我们必须利用遍历性这一性质：随机过程单次实现的时间平均（在足够长的时间段内）会收敛到所有可能实现的总体平均。假设 $r(t)$具有遍历性，随着数据集总时长增大（$T_{\text{N}} \to \infty$），经验时间平均误接受率几乎必然收敛到理论总体概率：

$$\lim_{T_{\text{N}} \to \infty} \widehat{P}_{\text{fa}}(\tau) = \lim_{T_{\text{N}} \to \infty} \frac{1}{T_{\text{N}}} \int_0^{T_{\text{N}}} 1(r(t) > \tau)\, dt \tag{49}$$

$$= \mathbb{E}_r\big[1(r > \tau)\big|H_{\text{N}}\big] \tag{50}$$

$$= \int_{\tau}^{+\infty} p(r|H_{\text{N}})\, dr. \tag{51}$$

同理，我们有

$$\lim_{T_{\text{Y}} \to \infty} \widehat{P}_{\text{miss}}(\tau) = \lim_{T_{\text{Y}} \to \infty} \frac{1}{T_{\text{Y}}} \int_0^{T_{\text{Y}}} 1(r(t) < \tau)\, dt \tag{52}$$

$$= \mathbb{E}_r\big[1(r < \tau)\big|H_{\text{Y}}\big] \tag{53}$$

$$= \int_{-\infty}^{\tau} p(r|H_{\text{Y}})\, dr. \tag{54}$$

因此，阈值 τ_{EER} 应满足

$$\lim_{T_{\text{Y}} \to \infty} \widehat{P}_{\text{miss}}(\tau_{\text{EER}}) = \int_{\tau_{\text{EER}}}^{+\infty} p(r|H_{\text{N}}) dr = \int_{-\infty}^{\tau_{\text{EER}}} p(r|H_{\text{Y}}) dr \tag{55}$$

$$= \lim_{T_{\text{N}} \to \infty} \widehat{P}_{\text{fa}}(\tau_{\text{EER}}). \tag{56}$$

[6]若要从每 N 毫秒仅输出一个分数的检测器中获取分数，在任意 t 时刻生成分数的任务类似于数模转换: https://en.wikipedia.org/wiki/ Digital-to-analog_converter。

[7]子区段的数量必须是整数，但为避免符号过于繁杂，我们省略向上取整或向下取整算子。

[8]https://en.wikipedia.org/wiki/Riemann_和

[9]原论文通过两次求和运算，将每个分数与每个真实标签片段进行比较。这只是为了找出重叠区域。

[10]目前，还没有理论能够解释为何应以这种方式定义风险。这仅仅是基于直觉：风险应当随时间进行积分。

因此，在较大的评估集上，求得 Range-EER 很可能与传统等错误率具有相同的解释：相对于贝叶斯决策阈值（或先验概率）的最大风险。

然而，上述推导要求满足

1. 平稳假设： $p(r(t)|H) \equiv p(r|H)$。2. 遍历性假设：随机过程的单个实现的时间平均值（$r(t)$）收敛于跨所有可能实现的总体平均值 $\mathbb{E}_{R(t)}$。3. 对数似然比 $r(t)$ 用于最优贝叶斯决策，且决策代价为 $C_{\text{fa}} = C_{\text{fr}} = 1$。

这些假设在实践中是否得到满足是另一个问题。**在实现中**。基于遍历性，我们可以根据公式（41–42）实现这两个错误率：

$$\widehat{R}(\pi_{\mathsf{Y}}, C_{\text{fa}}^{\mathsf{N}}, C_{\text{fr}}^{\mathsf{Y}}) \tag{57}$$

$$= C_{\text{fa}}^{\mathsf{N}}(1-\pi_{\mathsf{Y}})\widehat{P}_{\text{fa}}(\tau) + C_{\text{fr}}^{\mathsf{Y}}\pi_{\mathsf{Y}}\widehat{P}_{\text{miss}}(\tau) \tag{58}$$

$$= \frac{C_{\text{fa}}^{\mathsf{N}}(1-\pi_{\mathsf{Y}})}{T_{\mathsf{N}}} \int_0^{T_{\mathsf{N}}} 1(r(t) > \tau_{\text{Bayes}})\, dt \tag{59}$$

$$+ \frac{C_{\text{fr}}^{\mathsf{Y}}\pi_{\mathsf{Y}}}{T_{\mathsf{Y}}} \int_0^{T_{\mathsf{Y}}} 1(r(t) < \tau_{\text{Bayes}})\, dt \tag{60}$$

计算风险的具体步骤如下：

1. 给定一个大型评估集，我们将真实标签为伪造和真实的片段分别拼接成两个时长为T_{N} 和 T_{Y}的长话语。我们还将模型的输出$r(t)$进行拼接。2. 给定一个阈值 τ，按照公式（41–42）（或范围等错误率论文中的步骤）计算 $\widehat{P}_{\text{fa}}(\tau)$和 $\widehat{P}_{\text{miss}}(\tau)$。3. 将它们代入公式（60），并计算 $\widehat{R}(\pi_{\mathsf{Y}},\ C_{\text{fa}}^{\mathsf{N}},\ C_{\text{fr}}^{\mathsf{Y}})$。

A.3. 结论

范围等错误率与传统二分类评估相关，但能够持续测量整个时间段内的检测性能。在得分过程满足平稳性和遍历性的假设下，范围等错误率保留了传统等错误率的理论基础，并表示相对于贝叶斯决策阈值、采用均匀损失函数时的最大决策风险。

关于得分过程、决策代价等方面的假设，是为了建立范围等错误率与传统等错误率之间的联系而引入的。未来的工作需要重新思考这些假设是否必要或可靠。