利用 CTC 编码器草稿对基于大语言模型的自动语音识别进行自推测解码

George Saon、*Samuel Thomas*、*Takashi Fukuda*、*Tohru Nagano*、*Avihu Dekel*、*Luis Lastras*

IBM Research

gsaon@us.ibm.com

arXiv:2603.11243v1 [eess.AS] 11 Mar 2026

摘要

我们提出一种面向语音感知大语言模型的自推测解码方法，将连接时序分类编码器用作草稿模型，以加速自回归推理并提高自动语音识别准确率。该方法分为三个步骤：(1) 如果连接时序分类输出分布的帧熵低于阈值，则接受贪心 CTC 假设作为最终结果；(2) 否则，基于词元似然采用宽松接受准则，通过一次大语言模型前向传播验证 CTC 假设；(3) 如果验证失败，则从已接受的 CTC 前缀恢复自回归解码。在九个语料库和五种语言上进行的实验表明，该方法能够同时加速解码并降低词错误率。在 HuggingFace 开放自动语音识别基准上，使用一个拥有 10 亿参数的大语言模型和一个拥有 4.4 亿参数的 CTC 编码器，我们取得了创纪录的 5.58% 词错误率，并将实时因子倒数提升了 4.4 倍，而相较于自回归搜索，词错误率仅相对增加 12%。代码和模型权重以宽松许可证公开发布。

索引词：语音识别、语音感知大语言模型、自推测解码

1. 引言

作为注意力编码器-解码器（AED）模型的一种特殊形式 [1], , 语音感知语言模型（简称 SLM）目前是识别准确率最高的自动语音识别系统。事实上，HuggingFace Open ASR 排行榜 [2] 上排名前五的开源模型都是架构相似的 SLM。它们都使用 Conformer 声学编码器，并通过 RNN-T 损失 [3, 4], 、CTC 损失 [5], 或作为注意力编码器-解码器自动语音识别模型一部分的下一词元交叉熵损失 [6, 7]进行训练；语音模态适配器则采用多层感知机 [8, 6] 或查询变换器 [5]。适配器的作用是进行时间下采样，并将编码器输出的声学嵌入投影到大语言模型可理解的空间中。对于自动语音识别，大语言模型经过训练后会执行一种重复类任务 [9] ，也就是说，输出词元的嵌入可以视为语音嵌入的去噪版本。

AED 模型的一个主要缺点，以及由此延伸出的 SLMs 的一个主要缺点，是推理以自回归方式进行，每次生成一个词元，因此每生成一个词元，都需要通过文本大语言模型进行一次前向传播。与非自回归方法相比，这限制了并行度；在非自回归方法中，输出词元可同时预测，例如贪心解码 CTC、采用掩码预测纠错的 CTC [10] 或迭代重新对齐 [11]。另一类模型在准确率与推理速度之间实现了良好的权衡，称为半自回归模型；这类模型会在帧块内一次预测多个输出词元 [12] ，或联合预测词元和时长 [13]。

通过推测解码加速文本大语言模型的自回归推理 ve 是当前活跃的研究领域。在 [14] 和 [15], 中，作者使用一个小型草稿自回归模型，与大型目标模型并行运行，每次根据当前假设生成 K 个未来词元，再由目标模型通过一次前向传播进行验证。如果验证失败，目标模型便从草稿模型与目标模型分布出现差异的第一个位置开始回退到自回归解码。自投机解码的核心思想是将目标模型本身复用为草稿模型：可以像 [16]那样，通过 $i+1 \ldots i+N$ 使用 N 个独立的语言模型头同时预测未来位置的词元；也可以使用中间层的预测（提前退出） [17]。

文献 [18] 讨论了将推测解码应用于基于Transformer的编码器-解码器自动语音识别模型：作者将一个小型词元与时长转换器（TDT） [13] 解码器与冻结的编码器一同训练，后者作为较大因果Transformer解码器的草稿模型。该研究与我们提出的方法 [18] 有几处不同。首先，我们复用 SLM 的 CTC 编码器，而不是为草稿模型单独训练一个解码器。其次，作者着重于缓解重复错误，而我们展示了 CTC 与 SLM 之间互补的错误模式实际上能够降低词错误率。第三，我们使用由 CTC 输出分布的帧级熵给出的置信度指标 [19]，从而在 CTC假设置信度较高时，完全跳过 SLM 验证。

与我们工作相关的其他技术包括两遍式方法：第一遍使用 RNN-T 或自回归等快速模型，然后对候选假设进行重评分 [20], 或使用较慢的自回归模型优化输出 [21], ；以及在推理过程中进行分数融合的联合连接时序分类/注意力模型 [22]。

关于我们所提出方法的局限性，我们指出：(1) 该方法需要一个经过连接时序分类训练的编码器的小型语言模型，并且在投影器训练和大语言模型的 LoRA 微调期间，该编码器必须保持冻结；(2) 我们的技术仅适用于自动语音识别，不适用于翻译或口语问答等其他语音任务；(3) 小型语言模型对连接时序分类假设的验证是基于话语的，这意味着如果验证失败，就必须从验证失败的位置开始对整个话语进行自回归解码 [1] ，这限制了该方法在接受率较低的语料库上的收益。我们通过采用宽松接受标准来缓解问题 (3)：该标准要求连接时序分类假设在小型语言模型分布下是合理的，而不是与小型语言模型验证假设完全匹配。

本文结构如下：第 2 节介绍我们提出的方法，第 3 节展示该方法有效性的实验依据，第 4 节总结研究结果并提出未来的研究方向。

[1]虽然可以通过局部动态规划动态地重新接入连接时序分类草稿，但我们发现这种方法对于批量推理而言效率不高。

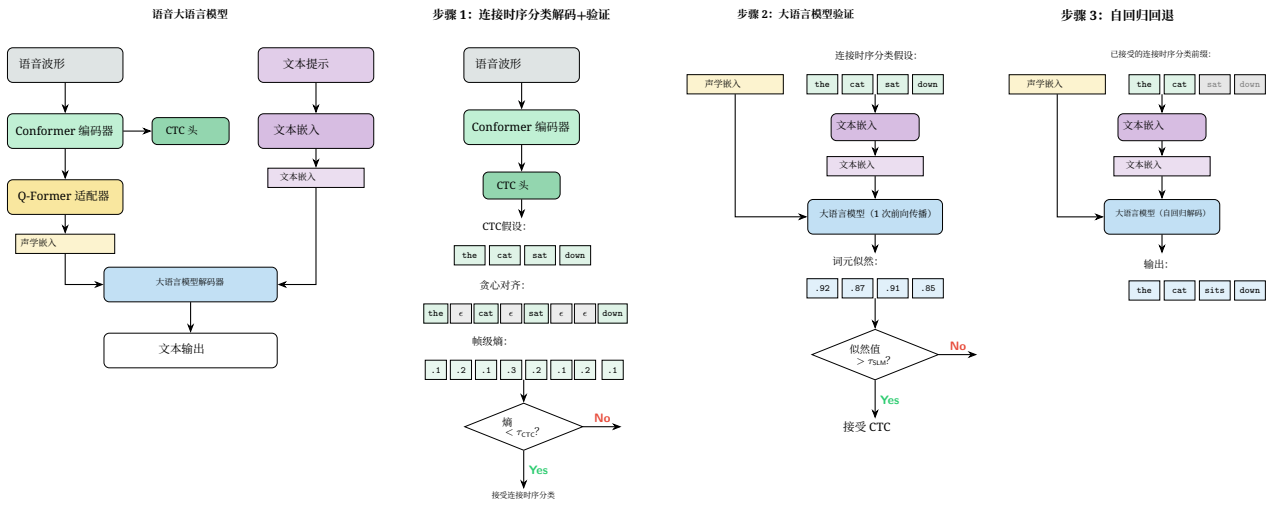


图 1： 面向语音感知大语言模型的自推测解码方法概览。

2. 方法描述

本节将详细介绍我们的方法，具体包括 CTC解码与验证过程（第 2.1 小节）、SLM验证过程（第 2.2 小节），以及 SLM自回归回退过程（第 2.3 小节）。这三个步骤以及所选用的 SLM 架构如图 1 所示。

2.1. CTC解码和验证

为引入一些记号，我们回顾一下，CTC公式 [23] 对输出序列 $\mathbf{y}|\mathbf{x}$ 的条件分布 p_{CTC} 进行建模，其中输出序列 $\mathbf{y} = (y_1, \ldots, y_U) \in \mathcal{Y}^*$ 长度为 U，给定长度为 $T \geq U$ 的输入序列 $\mathbf{x} = (x_1, \ldots, x_T) \in \mathcal{X}^*$。$\mathbf{x}$ 的元素是连续的多维向量，而 $\mathbf{y}$ 的元素则属于离散输出空间。 $p(\mathbf{y}|\mathbf{x})$ 表示为所有与 $\mathbf{y}$ 一致的可能对齐方式 $\mathbf{a} = (a_1, \ldots, a_T)$ 之和：

$$p_{CTC}(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{B}^{-1}(\mathbf{y})} p(\mathbf{a}|\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{B}^{-1}(\mathbf{y})} \prod_{t=1}^{T} p(a_t|h_t) \tag{1}$$

其中 $\mathbf{h} = (h_1, \ldots, h_T) = Encoder(\mathbf{x})$ 是由编码器计算得到的声学嵌入序列，$\mathbf{a}$ 中的元素属于扩展词汇表 $\mathcal{Y} = \mathcal{Y} \cup \{\epsilon\}$，其中 ϵ（称为空白符）表示空输出。 $\mathcal{B}$ 是一个映射，按此顺序移除连续重复项和空白符，使得 $\mathcal{B}(\mathbf{a}) = \mathbf{y}$。我们将草稿 CTC 假设计算为 $\hat{}\mathbf{y} = \mathcal{B}(\hat{}\mathbf{a})$，其中

$$\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_T) = (\operatorname*{argmax}_{a \in \overline{\mathcal{Y}}} p(a|h_1), \ldots, \operatorname*{argmax}_{a \in \overline{\mathcal{Y}}} p(a|h_T)) \tag{2}$$

表示贪心对齐路径。如果 CTC 输出分布的所有帧级熵都低于某个阈值，即

$$\sum_{a \in \overline{\mathcal{Y}}} -p(a|h_t) \log p(a|h_t) < \tau_{CTC} \qquad \forall t \in \{1, \ldots, T\} \tag{3}$$

2.2. 小型语言模型对 CTC假设的验证

与式（1）相比，语音语言模型以不同方式对给定输入序列 $\mathbf{y}$ 的输出序列$\mathbf{x}$的条件分布进行建模：

$$p_{SLM}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{U} p(y_i|\mathbf{y}_{<i}, \mathbf{x}) = \prod_{i=1}^{U} p(y_i|\mathbf{y}_{<i}, \mathbf{g}) \tag{4}$$

其中 $\mathbf{g} = (g_1, \ldots, g_{T'}) = Adapter(\mathbf{h})$，$T' \leq T$，是由语音模态适配器计算得到的、与大语言模型兼容的声学嵌入序列；该适配器以 CTC 编码器计算得到的嵌入序列 $\mathbf{h}$为输入。如果草稿 CTC 假设 $\hat{}\mathbf{y}$ 中所有词元在小型语言模型输出分布下的似然值都高于阈值，则该假设会在小型语言模型验证过程中被接受：

$$p(\hat{y}_i|\hat{\mathbf{y}}_{<i}, \mathbf{g}) > \tau_{SLM} \qquad \forall i \in \{1, \ldots, |\hat{\mathbf{y}}|\} \tag{5}$$

请注意，由于文本大语言模型的因果注意力掩码，(5) 可通过一次大语言模型前向传播对所有词元并行计算。我们在 (3) 中使用熵、在 (5) 中使用似然，是因为后者需要验证特定假设 $\hat{}\mathbf{y}$。

2.3. 带有 CTC 前缀的自回归回退传播

若 (5) 不成立，我们找出最长的已验证 CTC 前缀 $\hat{}\mathbf{y}_{<j}$，其中 $j = \min\{i \in \{1, \ldots, |\hat{}\mathbf{y}|\}\ s.t.\ p(\hat{y}_i|\hat{}\mathbf{y}_{<i}, \mathbf{g}) \leq \tau_{SLM}\}$，并直接从该处继续自回归生成：

$$\overline{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}^*} p_{SLM}(\mathbf{y}|\hat{\mathbf{y}}_{<j}, \mathbf{x}) \tag{6}$$

总而言之，自推测解码（简称 SSD）过程输出的最终假设是

$$\mathbf{y}_f = \begin{cases} \hat{\mathbf{y}}, & \text{if (3) or (5)} \\ (\hat{\mathbf{y}}_{<j}, \overline{\mathbf{y}}), & \text{otherwise} \end{cases} \tag{7}$$

表1： 比较 *granite-speech-4.0-1b* 的完全自回归与自推测解码。高准确率： τ_{CTC}=0.7、τ_{SLM}=0.2。高实时率倍数： τ_{CTC}=3.0、τ_{SLM}=0.1。*%C/%L* = 在每个验证阶段的连接时序分类/大语言模型接受率。实时因子倒数均在 *1* 个 *H100* 上测得。 $^{*}p < 0.0001$、$^{\dagger}p < 0.05$ （统计显著性采用双比例 *z* 检验进行评估）。

	完全自回归		高准确率				高实时率倍数	
测试集	词错误率	实时因子倒数	词错误率	实时因子倒数	%C	%L	词错误率	实时因子倒数
英语（*Open ASR*）								
AMI-IHM	8.65	455	8.31	621	39	39	9.40	1716
Earnings22	9.19	391	8.96	278	6	19	11.44	1982
GigaSpeech	10.19	511	9.95	514	10	51	10.67	2503
LS Clean	1.37	510	1.37	712	17	73	1.63	2160
LS Other	3.02	476	2.88	667	18	61	3.77	2108
SPGISpeech	3.97	666	3.80	592	6	44	4.48	2736
TED-LIUM	3.38	358	3.35	444	9	60	3.91	1748
VoxPopuli	6.21	304	5.99	365	12	51	7.18	2035
Avg.	**5.75**	**564**	**5.58***	**548**			**6.56**	
多语言（*MLS*）								
MLS 英语	4.74	702	4.72	772	1	42	5.75	2983
MLS 德语	4.59	663	4.42	755	3	48	4.91	2755
MLS 西班牙语	3.28	627	3.22	779	3	50	3.66	2490
MLS 法语	4.51	634	4.48	658	1	39	5.52	2834
MLS 葡萄牙语	10.25	521	9.84	531	8	10	8.83	2701
Avg.	**5.47**	**629**	**5.33**†	**699**	3	38	**5.73**	**2753**
多语言（*CommonVoice*）								
CV 英语	6.52	954	6.56	716	28	31	9.30	2446
CV 德语	4.79	872	4.71	998	0	59	6.34	2309
CV 西班牙语	4.04	752	3.98	890	1	59	5.42	2526
CV 法语	7.15	674	7.10	910	0	46	10.69	2341
CV 葡萄牙语	2.82	866	2.80	1066	5	69	3.83	2344
Avg.	**5.06**	**824**	**5.03**	**916**	7	53	**7.11**	

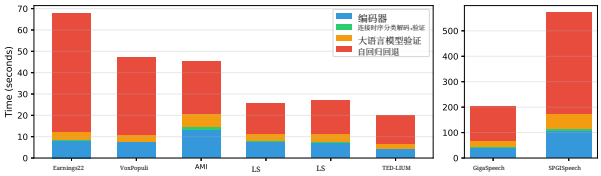


图 2： *granite-speech-4.0-1b* 在高准确率 *SSD* 区域中各个处理阶段的运行时间和细分。

3. 实验与结果

我们选择 Granite Speech [5]作为实验中的小型语言模型（SLM），因为它满足以下约束：使用通过连接时序分类（CTC）训练且已冻结的声学编码器，并且在 Open ASR 排行榜 [2]上表现出色。我们报告了现有 granite-speech-3.3-2b 和 granite-speech-3.3-8b 模型的结果，以及在拥有 1B 参数的 granite-4.0-nano 大语言模型（LLM）之上新训练的小型语言模型（SLM）的结果 [24]。对于后者，我们将从训练和评估数据、编码器架构及 CTC 训练，到投影器架构及结合大语言模型微调的训练，讨论其架构和训练细节。

3.1. 数据、架构和训练细节

我们的模型使用来自英语、法语、德语、西班牙语、葡萄牙语和日语主要公开多语言自动语音识别数据集的 90,000 小时数据进行训练，还使用了 CommonVoice 与英语之间的双向合成翻译数据，以支持语音翻译任务。具体而言，

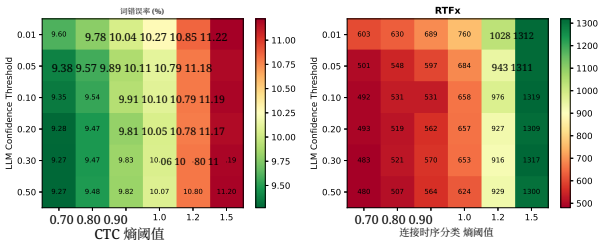


图3： *Earnings-22* 上接受阈值 τ_{CTC} 和 τ_{SLM} 对词错误率和实时因子的倒数的影响。

我们的 CTC 编码器使用多语言 LibriSpeech [25], CommonVoice 17.0 [26], LibriSpeech [27], Voxpopuli [28],AMI [29], YODAS [30], Earnings-22 [31]（训练划分，参见 [32]）、Switchboard [33], CallHome、Fisher [34], Voice-mail [35] 和 ReazonSpeech [36]进行训练。投影器训练和大语言模型微调使用前述语料库，以及使用 Phi-4 [37], MAD-LAD [38] 和 Granite 3 [39] 翻译生成的 21,000 小时合成翻译数据。我们还报告了在 Gigaspeech [40], SPGI Speech [41] 和 TED LIUM [42]上的自动语音识别结果；由于这些数据集采用非商业许可，因此未纳入训练数据。

这个拥有 4.4 亿参数的编码器由 16 个隐藏维度为 1024 的 Conformer 层组成，采用 4 秒块自注意力，并以中间层的预测结果作为条件。大小为 348 的输出层包含前 256 个 ASCII 字符条目，以覆盖欧洲语言，另有 92 个用于日语的片假名字符。它在此前指定的所有自动语音识别语料库上，使用字符级连接时序分类训练了 20 个周期，批量大小为 256，并结合均衡采样和自然分布采样。

我们使用一个两层、拥有 3700 万参数的查询变换器（q-former）投影器，对每 15 个连续帧窗口中的声学嵌入进行下采样，下采样因子为 5；每层使用 3 个可训练查询 [43]。初始的 10 毫秒、80 维对数梅尔帧在时间维度上按 10 倍下采样（编码器下采样 2 倍，投影器下采样 5 倍）。投影器和用于大语言模型线性模块的秩为 64 的 LoRA 适配器在此前提及的所有语料库上联合训练 3 个周期，共进行 90 万次更新，峰值学习率为 1e-4，批量大小为 128 个语句，并采用均衡采样。

3.2. 实验细节

在所有实验中，我们使用 bfloat16 精度，在 1 个 H100 上以批量解码方式测量 WER 和实时因子的倒数（RTFx = 音频时长除以处理时间）。为充分利用图形处理器，我们按音频时长对语句排序，并使用由最大词元数控制的自适应批处理（实际设置为 50K）；对于连接时序分类+大语言模型验证阶段和自回归回退阶段，我们分别进行批处理。对于在各处理阶段之间验证失败的语句，我们会将其声学嵌入卸载至 CPU。其他优化包括：(1) 合并低秩适配参数与大语言模型权重，以避免低秩适配计算开销；(2) 为 Granite-Speech 单独实例化大语言模型，以启用闪存注意力（FA2），因为 Granite-Speech 原生不支持 FA2；(3) 仅在连接时序分类验证位置计算大语言模型逻辑值，而不是对包含文本指令和声学嵌入的整个提示进行计算。

在图3中，我们研究了连接时序分类和大语言模型验证阈值对 Earnings-22 语料库的词错误率和实时因子倒数的影响

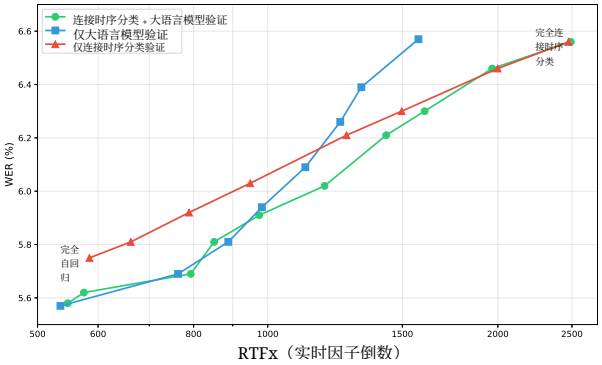


图4： 验证阶段消融研究：绿色 =表示先经过连接时序分类，再经过大语言模型；蓝色 = 表示大语言模型验证；红色 = 表示连接时序分类验证。结果基于 *ESB/Open ASR* 测试集报告。

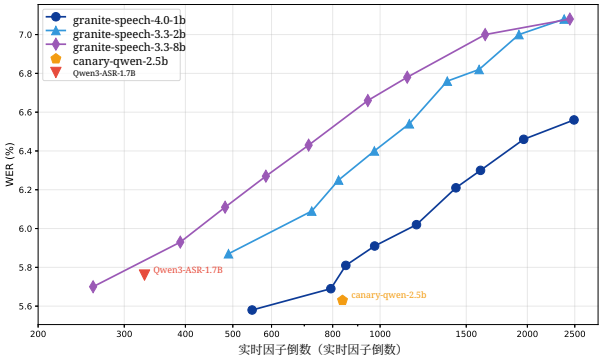


图5：*Granite-Speech*与领先的小语言模型在*Open ASR*测试集上的词错误率和实时因子倒数对比（均在*1*块*H100*上）。

语料库，并得出结论： τ_{CTC} 对工作点的影响要大得多。因此，在所有实验中，我们固定 $\tau_{SLM} = 0.1$ ，仅改变 $\tau_{CTC} \in [0.7, 3.0]$ ；表1中的高准确率设置除外，在这些设置中 $\tau_{SLM} = 0.2$。

3.3. 验证阶段消融研究

本节探讨的问题是：在整个词错误率-逆实时因子范围内，两个验证阶段是否都是实现最佳性能所必需的。具体而言，我们比较同时使用连接时序分类和大语言模型接受阶段，与仅使用一个验证阶段（连接时序分类或大语言模型）的效果。图4的结果表明，停用连接时序分类接受阶段（即直接将连接时序分类假设送入大语言模型验证），在高准确率区间的性能与基线相近，但在高吞吐量端的实时因子倒数显著更低。相反，停用大语言模型验证阶段（即仅使用连接时序分类熵接受机制，并对未接受的样本采用完全自回归回退），无法在高准确率区间达到最低词错误率（5.75%，而非5.58%），但在实时因子倒数较高时，其曲线与基线相当。所提出的同时采用两个验证步骤的方法在大部分范围内占优，并实现了最佳帕累托前沿。

3.4. 大语言模型验证比完全自回归的准确率更高

从表1和图4的结果可以看出，与完全自回归解码相比，对连接时序分类假设进行大语言模型验证能够降低所有语料库上的词错误率。我们推测，这是连接时序分类与小型语言模型之间的系统组合效应所致，因为两者产生的识别错误具有互补性。表2展示了一些示例，其中经过大语言模型验证的连接时序分类假设比完全自回归解码得到的结果更准确。显然，自回归输出往往更加流畅，但并不总是忠实于声学信号。这是 Whisper 等编码器-解码器模型的一个已知问题，称为语言模型偏差 [44]。

3.5. 详细结果

在表1中，我们报告了 granite-speech- 4.0-1b 在 ESB/Open ASR 英语测试集以及两个多语言语料库（CommonVoice 17.0 和 MLS）上的词错误率和实时因子倒数结果。我们观察到，在高准确率区间，SSD 在所有语料库上的表现都优于完全自回归，且吞吐量没有损失。准确率的提升源于较低的 CTC 接受率（%C 列）以及较高的大语言模型接受率：CTC 假设中有 40-50% 被接受（%L

表2： 经大语言模型验证的 *CTC* 假设比完全自回归解码更准确的示例。错误以粗体显示。

AMI
REF：所以如果你能处理这件事，然后..... 比如，不，其实不是 CTC：所以如果你能处理这件事，然后，比如，不，其实不是 自回归：所以如果你能处理** 这件事，然后**我** 比如，不，其实不是
LibriSpeech test-other
REF：而我如同新郎 连接时序分类：而我如同 sponsis 自回归：而我**如**同新**郎**
REF：是什么把他从那边甩了下来 连接时序分类：什么**甩**了他，从那边 自回归：什么**trucked****em** off yonder
VoxPopuli
REF：商业 正在要求工人，他们正在要求穿着要求做更多 连接时序分类：商业正在要求工人正在要求穿着要求做更多 自回归：商业**es are** 正在要求工人正在要求穿着要求做更多
REF：我声明 对此事特别感兴趣 CTC：我声明 对此事特别感兴趣 自回归：我对此事特别感兴趣

列）。在高实时率倍数区间，通过将 CTC门控设为接近 100 %，我们在 Open ASR 上将推理速度提升了 4.4 倍，同时词错误率下降了 12%。至于运行时，图 2 显示，在高准确率区间，编码器和回退自回归阶段最耗时。

最后，在图5中，我们将三种 Granite-Speech 模型（3.3-2b、3.3-8b 和 4.0-1b）与 SSD 进行比较，并与最接近的两个竞争模型（canary-qwen-2.5b [8, 4] 和 qwen3-asr-1.7b [7]）对照。为确保实时因子倒数比较公平，我们使用这些模型官方提供的 HuggingFace 评测工具，自行对后两者运行了 Open ASR 基准测试。

4. 结论与未来工作

我们已经证明，使用快速、以非自回归方式计算的连接时序分类编码器草稿，可以提升语音感知大语言模型的准确率和推理速度。通过大语言模型验证基于声学信息的连接时序分类假设，可以减轻语言建模偏差导致的错误，从而提高准确率。通过接受高置信度连接时序分类草稿，可以提升吞吐量。重要的是，这一方法无需重新训练小型语言模型，也无需提供单独的草稿模型，只需利用编码器的连接时序分类头进行快速转录和门控，并由大语言模型负责验证和回退自回归解码，即可实现上述效果。未来的工作将探索如何专门针对推测解码（在保持准确率的同时实现较高的大语言模型接受率）联合训练编码器和大语言模型，并研究如何应用其中的一些思路来降低实时对话应用的延迟。

5. 生成式人工智能使用声明

实验中使用的代码是在编程助手 Claude（由 Anthropic 开发）的帮助下编写的。实验由人工运行，结果也由人工验证。生成式人工智能还用于表格排版和图表设计。论文其余部分均由人工撰写，未使用人工智能进行校对或编辑。作者对本投稿的内容承担全部责任。

6. 参考文献

[1] W. Chan、N. Jaitly、Q. Le 和 O. Vinyals，“Listen, attend and spell”，载于*ICASSP*，2016年。[2] V. Srivastav、S. Zheng、E. Bezzam、E. L. Bihan、A. Moumen 和 S. Gandhi，“Open ASR 排行榜：迈向可复现且透明的多语言语音识别评估”，*arXiv* 预印本 *arXiv:2510.06961*，2025年。[3] D. Rekesh、N. R. Koluguri、S. Kriman、S. Majumdar、V. Noroozi、H. Huang、O. Hrinchuk、K. Puvvada、A. Kumar、J. Balam等，“采用线性可扩展注意力的快速 Conformer，实现高效语音识别”，载于*2023*年*IEEE*自动语音识别与理解研讨会（*ASRU*）。IEEE，2023年，第1–8页。[4] M. Sekoyan、N. R. Koluguri、N. Tadevosyan、P. Zelasko、T. Bart-ley、N. Karpov、J. Balam 和 B. Ginsburg，“Canary-1b-v2 与 parakeet-tdt-0.6 b-v3：用于多语言自动语音识别与自动语音翻译的高效高性能模型”，*arXiv* 预印本 *arXiv:2509.14128*，2025年。[5] G. Saon、A. Dekel、A. Brooks、T. Nagano、A. Daniels、A. Satt、A. Mittal、B. Kingsbury、D. Haws、E. Morais等，“Granite-speech：具备强大英语自动语音识别能力的开源语音感知大语言模型”，*arXiv* 预印本 *arXiv:2505.08699*，2025年。[6] A. Abouelenin、A. Ashfaq、A. Atkinson、H. Awadalla、N. Bach、J. Bao、A. Benhaim、M. Cai、V. Chaudhary、C. Chen等，“Phi-4-mini 技术报告：通过 LoRA 混合实现紧凑而强大的多模态语言模型”，*arXiv* 预印本 *arXiv:2503.01743*，2025年。[7] X. Shi、X. Wang、Z. Guo、Y. Wang、P. Zhang、X. Zhang、Z. Guo、H. Hao、Y. Xi、B. Yang、J. Xu、J. Zhou 和 J. Lin，“Qwen3-ASR 技术报告”，2026年。 [在线]。可访问：https://arxiv.org/abs/2601.21337[8] Z. Chen、H. Huang、A. Andrusenko、O. Hrinchuk、K. C. Puvvada、J. Li、S. Ghosh、J. Balam 和 B. Ginsburg，“Salm：用于语音识别与翻译的上下文学习语音增强语言模型”，载于*ICASSP 2024-2024 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024年，第13 521–13 525页。[9] A. Grattafiori、A. Dubey、A. Jauhri、A. Pandey、A. Kadian、A. Al-Dahle、A. Letman、A. Mathur、A. Schelten、A. Vaughan等，“Llama 3*21783*，2024*2407*模型家族”，*arXiv* 预印本 *arXiv:* 。[10] Y. Higuchi、S. Watanabe、N. Chen、T. Ogawa 和 T. Kobayashi，“Mask CTC：结合连接时序分类与掩码预测的非自回归端到端自动语音识别”，*arXiv* 预印本 *arXiv:2005.08700*，2020年。[11] E. A. Chi、J. Salazar 和 K. Kirchhoff，“Align-Refine：通过迭代式重新对齐实现非自回归语音识别”，载于*2021*年北美计算语言学协会分会年会论文集：人类语言技术，2021年，第1920–1927页。[12] S. Arora、G. Saon、S. Watanabe 和 B. Kingsbury，“结合标签上下文的半自回归流式自动语音识别”，载于*ICASSP 2024- 2024 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024年，第11 681–11 685页。[13] H. Xu、F. Jia、S. Majumdar、H. Huang、S. Watanabe 和 B. Gins-burg，“通过联合预测词元和时长实现高效序列转换”，载于国际机器学习会议。PMLR，2023年，第38 462–38 484页。[14] Y. Leviathan、M. Kalman 和 Y. Matias，“通过推测解码实现 Transformer 快速推理”，载于国际机器学习会议。PMLR，2023年，第19 274–19 286页。[15] C. Chen、S. Borgeaud、G. Irving、J.-B. Lespiau、L. Sifre 和 J. Jumper，《通过推测采样加速大型语言模型解码》，*arXiv* 预印本 *arXiv:2302.01318*，2023 年。[16] T. Cai 、 Y. Li 、 Z. Geng 、 H. Peng、J. D. Lee、D. Chen 和 T. Dao，《Medusa：一种采用多个解码头的简单大语言模型推理加速框架》，*arXiv* 预印本 *arXiv:2401.10774*，2024年。[17] J. Zhang、J. Wang、H. Li、L. Shou、K. Chen、G. Chen 和 S. Mehrotra，《Draft& verify：通过自推测解码实现无损大型语言模型加速》，载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），2024年，第11 263–11 282页。[18] Y. Lim、J. Park、H. Y. Kim、H. Lee 和 B.-Y. Kim，《混合解码：序列模型的快速通道与选择性细致纠正》，*arXiv* 预印本 *arXiv:2508.19671*，2025年。[19] A. Laptev 和 B. Ginsburg，《端到端自动语音识别中快速的基于熵的词级置信度估计方法》，载于*2022*年*IEEE*口语技术研讨会（*SLT*）。IEEE，2023年，第152–159页。[20] T. N. Sainath、R. Pang、D. Rybach、Y. He、R. Prabhavalkar、W. Li、M. Visontai、Q. Liang、T. Strohman、Y. Wu 等，《两遍式端到端语音识别》，*arXiv* 预印本 *arXiv:1908 10992*，2019。。[21] K. Hu、T. N. Sainath、R. Pang 和 R. Prabhavalkar，《基于审议模型的两遍式端到端语音识别》，载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年，第7799–7803页。[22] S. Watanabe、T. Hori、S. Kim、J. R. Hershey 和 T. Hayashi，《用于端到端语音识别的混合连接时序分类/注意力架构》，*IEEE*信号处理精选主题期刊，第11卷，第8期，第1240–1253页，2017年。[23] A. Graves、S. Fern´a ndez、F. Gomez 和 J. Schmidhuber，《连接时序分类：使用循环神经网络对未分段序列数据进行标注》，载于第*23*届国际机器学习会议论文集，2006年，第369–376页。[24] IBM Research，《Granite 4.0语言模型》，https://github.com/ibm-granite/granite-4.0-language-models，2025年，访问日期：2025-10-01。[25] V. Pratap、Q. Xu、A. Sriram、G. Synnaeve 和 R. Collobert，《MLS：用于语音研究的大规模多语言数据集》，*arXiv* 预印本 *arXiv:2012.03411*，2020年。[26] R. Ardila、M. Branson、K. Davis、M. Henretty、M. Kohler、J. Meyer、R. Morais、L. Saunders、F. M. Tyers 和 G. Weber，《Common Voice：大规模多语言语音语料库》，*arXiv* 预印本 *arXiv:1912.06670*，2019年。[27] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，《LibriSpeech：基于公有领域有声读物的自动语音识别语料库》，载于*2015*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2015年，第5206–5210页。[28] C. Wang、M. Riviere、A. Lee、A. Wu、C. Talnikar、D. Haz-iza、M. Williamson、J. Pino 和 E. Dupoux，《VoxPopuli：用于表征学习、半监督学习和解释的大规模多语言语音语料库》，*arXiv* 预印本 *arXiv:2101.00390*，2021年。[29] W. Kraaij、T. Hain、M. Lincoln 和 W. Post，《AMI会议语料库》，载于行为研究方法与技术国际会议论文集，2005年，第1–4页。[30] X. Li、S. Takamichi、T. Saeki、W. Chen、S. Shiota 和 S. Watan-abe，《YODAS：面向YouTube的音频与语音数据集》，载于*2023*年*IEEE*自动语音识别与理解研讨会（*ASRU*）。IEEE，2023年，第1–8页。[31] M. D. Rio、P. Ha、Q. McNamara、C. Miller 和 S. Chandra，《Earnings-22：真实场景中的口音实用基准测试》，2022年。[32] S. Gandhi、P. von Platen 和 A. M. Rush，《ESB：多领域端到端语音识别基准测试》，2022年。 [在线]。可访问：https://arxiv.org/abs/2210.13352

[33] J. J. Godfrey、E. C. Holliman 和 J. McDaniel，《Switchboard：用于研究与开发的电话语音语料库》，载于*IEEE*声学、语音与信号处理国际会议，第1卷。IEEE计算机学会，1992年，第517–520页。[34] C. Cieri、D. Miller 和 K. Walker，《Fisher语料库：面向新一代语音转文本研究的资源》，载于语言资源与评测会议，第4卷，2004年，第69–71页。[35] M. Padmanabhan、G. Saon、J. Huang、B. Kingsbury 和 L. Mangu，《语音邮件转录任务中的自动语音识别性能》，*IEEE*语音与音频处理汇刊，第10卷，第7期，第433–442页，2002年。[36] Y. Yin、D. Mori 和 S. Fujimoto，《ReazonSpeech：用于日语自动语音识别的免费大型语料库》，载于第*29*届自然语言处理年会论文集（*NLP 2023*），2023年。[37] M. Abdin、J. Aneja、H. Behl、S. Bubeck、R. Eldan、S. Gunasekar、M. Harrison、R. J. Hewett、M. Javaheripi、P. Kauffmann 等，《Phi-4技术报告》，*arXiv* 预印本 *arXiv:2412.08905*，2024年。[38] S. Kudugunta、I. Caswell、B. Zhang、X. Garcia、D. Xin、A. Kusu-pati、R. Stella、A. Bapna 和 O. Firat，《MADLAD-400：一个多语言、文档级且经过审计的大型数据集》，神经信息处理系统进展，第36卷，67 296，2023年 67 284页。– 。[39] IBM Research，《Granite 3.0语言模型》，https://github.com/ibm-granite/granite-3.0-language-models，2024年。[40] G. Chen、S. Chai、G. Wang、J. Du、W.-Q. Zhang、C. Weng、D. Su、D. Povey、J. Trmal、J. Zhang 等，《GigaSpeech：一个不断发展的多领域自动语音识别语料库，包含10,000小时转录音频》，*arXiv* 预印本 *arXiv:2106.06909*，2021年。[41] P. K. O’Neill、V. Lavrukhin、S. Majumdar、V. Noroozi、Y. Zhang、O. Kuchaiev、J. Balam、Y. Dovzhenko、K. Freyberg、M. D. Shul-man 等，《SPGISpeech：用于全格式端到端语音识别的5,000小时金融音频转录语料》，*arXiv* 预印本 *arXiv:2104.02014*，2021年。[42] A. Rousseau、P. Del´eglise 和 Y. Esteve，《TED-LIUM：一个专用于自动语音识别的语料库》，载于语言资源与评测会议，2012年，第125–129页。[43] C. Tang、W. Yu、G. Sun、X. Chen、T. Tan、W. Li、L. Lu、Z. Ma 和 C. Zhang，《SALMONN：探索大型语言模型的通用听觉能力》，*arXiv* 预印本 *arXiv:2310.13289*，2023年。[44] A. Koenecke、A. S. G. Choi、K. X. Mei、H. Schellmann 和 M. Sloane，《Careless Whisper：语音转文本幻觉危害》，载于*2024*年*ACM*公平性、问责制与透明度会议论文集，2024年，第1672–1681页。