arXiv:2512.14652v1 [eess.AS] 16 Dec 2025

Se基于长篇声学编码的分段注意力解码 s

Pawel Sw ietojanski[1], *Xinwei Li*[1], *Mingbin Xu*[1],*Takaaki Hori*[1],*Dogan Can*[1],*Xiaodan Zhuang* [1]

[1]Apple，美国{pswietojanski, xinwei li2}
@apple.com

摘要

我们探讨了注意力式编码器—解码器（AED）模型与长篇声学编码表示之间的根本不兼容性。在分段语句上训练的AED 模型会利用分段边界之外有限的声学上下文，学会编码绝对帧位置；但在解码长篇片段时，这些线索会消失，模型因而无法泛化。由于交叉注意力中的键和值具有排列不变性，模型失去了对声学编码排序的能力。我们提出了四项改进：（1）为每个解码片段在交叉注意力中注入显式的绝对位置编码；（2）通过扩展声学上下文进行长篇训练，以消除隐式绝对位置编码；（3）拼接片段，以涵盖训练所需的多种分段方式；（4）采用语义分割，使 AED 解码的片段与训练片段对齐。我们证明，这些改进弥合了连续声学编码与分段声学编码之间的准确率差距，从而能够以自回归方式使用注意力解码器。**索引词**：注意力解码器，长篇自动语音识别

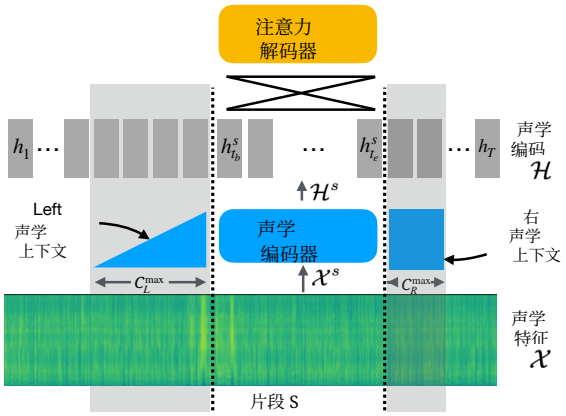

图 1： *LFE*、左/右声学上下文和边界条件示意图。阴影区域表示对 *LFE* 至关重要的区域，在分段数据的训练和推理过程中缺失。

1. 引言与相关工作

注意力编码器-解码器（AED）模型 [1, 2] 是端到端语音识别模型中一种流行的架构，具有强大的自回归能力，但基于 AED 模型的长篇（LF）自动语音识别仍然面临挑战 [3–5]。当前方法通常依赖外部分段系统或不同的窗口化技术 [6–8] ，以人为方式重现边界条件。因此，注意力解码器（AD）通常用于多遍设置：第一遍生成的假设由 AD 进行修正 [9,10]。然而，这种重打分作用于 N-best 列表或离散词元，而非完整的自回归解码，因此限制了准确率提升的潜力。近期与 LF 自动语音识别相关的研究大多聚焦于采用线性化注意力的 Transformer 编码器 [11, 12], 或替代架构 [13]。 [14]最近提出了长篇（LF）训练，并评估了不同类型的线性时间编码器在 LF 任务上的表现。在本研究中，我们关注 AD 无法以自回归方式使用 LF 编码的问题。此前已有研究指出，AD 在 LF 数据上的问题
此前据我们所知，尽管这一问题截至目前尚未得到解决。

我们的贡献主要有四点：（1）我们发现了长格式音频的基于注意力的自动语音识别中的置换不变性问题，并展示了短格式边界效应如何在训练期间掩盖这一局限。（2）为解决（1）中的问题，我们提出了多项改进，包括交叉注意力位置编码和长格式训练。（3）我们证明，我们的方法在分段和长篇场景下的性能相当，同时与规模相近的模型相比仍保持竞争力。（4）为模型扩展了首轮语义分割能力，并展示了其性能优于语音活动检测（VAD）。

2. 定义

2.1. 基于注意力的编码器-解码器模型

AED 模型通过三个主要部分将语音音频转换为文本：声学编码器、交叉注意力机制和自回归解码器 [1, 2]。

声学编码器： 编码器 $\mathcal{H} = \mathcal{E}_\theta(\mathcal{X})$ 接收一系列声学特征 $\mathcal{X} = (x_1, ..., x_{T'})$，并将其转换为一系列声学编码 $\mathcal{H} = (h_1, ..., h_T)$，其中 $T \leq T'$。现代编码器采用带有卷积下采样的 Conformer [15, 16] 或 Transformer [17, 18] 模块。

注意力解码器： 解码器 $y_u = \mathcal{D}_\theta(y_{u-1}, \mathcal{C}_{<u}, \mathcal{H})$ 以自回归方式每次生成一个输出词元 $\mathcal{Y} = (y_1, ..., y_U)$。在每个步骤 u，解码器会结合三种信息：前一个词元嵌入 y_{u-1}、截至 u 的前一个内部状态 $\mathcal{C}_{<u}$ ，以及声学编码 $\mathcal{H}$。

交叉注意力机制： $\mathcal{D}_\theta$ 通过交叉注意力 [17] 层与编码器$\mathcal{E}_\theta$ 交互。在 AED 中，点积交叉注意力的查询来自解码器状态 $\mathcal{C}$，而键和值则由声学编码 $\mathcal{H}$生成。

2.2. 长篇声学编码

在本文中，我们研究了基于 Transformer 的编码器 $\mathcal{E}_\theta$ 的感受野与注意力解码器 $\mathcal{D}_\theta$ 转录任意片段 $\mathcal{H}^s \in \mathcal{H}$的能力之间的关系，其中编码 $\mathcal{H}$ 由任意音频流 $\mathcal{X}$计算得到。与循环神经网络不同，Transformer 架构能够精确控制每一步所使用的邻域信息。例如，时间 t处的声学编码$h_t = \mathcal{E}_\theta(\mathcal{X})_t$，可能源自特征 $\mathcal{E}_\theta$ 中的任意子集，具体取决于

掩码 $\mathcal{M}$ 的结构，该掩码用于 $\mathcal{E}_\theta$的自注意力层。定义 1 和定义 2 给出了本文所涉及的长篇声学编码和长篇片段的概念:

定义1 (长篇声学编码（LFE）)。 当声学编码 h_t 满足以下条件时，它就是 *LFE*:

$$h_t = \mathcal{E}_\theta(\mathcal{X})_t \quad with \quad C_L(t) = C_L^{\max}, \quad C_R(t) = C_R^{\max}$$

其中，$C_L^{\max}$和 $C_R^{\max}$分别表示编码器 $\mathcal{E}_\theta$ 在处理输入 $\mathcal{X}$时所支持的最大左侧/右侧上下文。

定义 2 (长篇片段)。 如果片段 s 中的每个编码 $h_t^s \in \mathcal{H}^s$ 都是 *LFE* 编码，则该片段被视为长篇片段。

换言之，当编码 h_t 根据 $\mathcal{E}_\theta$的掩码策略，利用完整的双向上下文计算时，它就成为 LFE。图 1 展示了任意片段 S的短篇编码（SFE）与 LFE 之间的关键区别，以及其对应的声学编码 $\mathcal{H}^s = (h_{t_b}^s, \ldots\ldots, h_{t_e}^s)$。具体而言，对于分段语句（不含阴影区域） $C_L(t_b)$， $h_{t_b}^s$ 的值以及 $C_R(t_e)$， $h_{t_e}^s$ 的值均为 0。相比之下，对于 LFE（含阴影区域）， $C_L(t_b) = C_L^{\max}$，且 $C_R(t_e) = C_R^{\max}$。全局注意力是一个特殊情况，因为任意时刻的任何编码都可以相对于录音边缘处的编码（h_1 和 h_T）进行锚定，从而缓解长篇注意力解码问题。然而，全局注意力并非总是可行：i）编码器可能需要进行流式处理；更重要的是，ii）众所周知，语音全局注意力对训练条件与推理条件之间的时长不匹配很敏感 [19, 20], ；在长篇场景中，随着左侧上下文不断增长，这一问题可能会进一步加剧。一般而言，对于任意长度的序列，消除短篇/长篇不匹配并非易事，因为训练语句的长度通常比推理时未分段音频流的长度更受限制。

在这项工作中，我们将 $\mathcal{E}_\theta$ 的左侧和右侧上下文固定为一定数量的回看和前瞻 Conformer 帧，类似于 [18], ，以确保在此机制下进行稳健估计。根据这一表述，获得 LFE 所需的预期上下文取决于以下因素：每个自注意力层在前一层中能够看到的有限数量的 N “回看”帧、Transformer 层数 L 以及 R $C_L^{\max} = L \cdot N \cdot R$ 抽取因子 : 声学帧数。对于右侧上下文，我们采用因果分块 [21] , 以在延迟、准确率和硬件利用率之间取得良好平衡，因此 $C_R^{\max} = M \cdot R$，其中 M 是自注意力层在未来能够查看的分块内帧数上限。此外，与 [22] 类似，我们使用多种块大小训练模型，以便在推理期间配置延迟与准确率之间的权衡。这种上下文受限架构适用于长篇应用。

3. 提出的修改

3.1. 交叉注意力中的显式位置编码

点积交叉注意力的基本属性是其排列不变性 [17]。对于输入的任意顺序 $\mathcal{H}$，注意力都会产生相同的输出，这意味着仅凭内容无法区分不同的声学顺序，除非内容本身编码了序列顺序。

对于连续的长篇声学流，这一特性会带来问题，因为 AED 通常是在分段语句上训练的，而注意力可以隐式地学会利用边界效应作为位置锚点；在片段边界处，声学编码缺乏完整的双向上下文（参见 图 1 中未着色的区域）。这会形成鲜明的模式，有助于解码器隐式跟踪 $\mathcal{H}$的顺序。然而，当编码器处理足够长的声学信号时（如长语音场景所示，参见 图 1 中着色的区域），这些边界线索便会消失，所有嵌入在上下文上也显得相似，从而引发位置跟踪问题。这一局限会导致转录输出重复，并使注意力解码（AD）无法发出句末（EOS）标记。

为了解决排列不变性问题，我们为每个片段 $\mathcal{H}^s$添加绝对位置编码 $\mathbf{p}$，即 $\mathcal{H}_p^s = \mathcal{H}^s + \mathbf{p}$。这是在通常注入编码器输入 [17, 23], 的位置编码之外的补充，因为我们还会在 LFE 级别 $\mathcal{H}$显式加入这些编码，然后再将其馈入交叉注意力的键和值。这样可以确保注意力权重反映时间位置，同时与片段 s 相关的位置信息也会传递至解码器。关键的是，每个解码片段的位置索引都会重置，而不是沿用全局位置，从而在避免无界增长的同时提供充分的时间定位。

3.2. 数据级转换

通常，自动语音识别使用分段数据进行训练，这使模型可能（无意中）利用边界效应，而这些效应在长篇场景中可能不存在（见第 2.2 节）。为鼓励模型利用位置编码（第 3.1 节）中的显式信号来确定声学顺序，而不是依赖这些伪影，我们提出一种长语音训练方法，让注意力解码接触长篇片段（定义 2）。这不同于其他数据级长篇训练方案 [14,24], ，因为我们训练注意力解码，使其专门解码长篇音频编码流 $\mathcal{H}$中的长语音示例片段 $\mathcal{H}^s$ 。

扩展声学上下文（AC）： 在使用一对 $\{\mathcal{Y}^s, \mathcal{X}^s\}$ 进行训练时，我们将随机选取的长语音示例的声学编码扩展至 $\mathcal{X}_E^s = [\mathcal{X}_L^s, \mathcal{X}^s, \mathcal{X}_R^s]$，以纳入有效长语音片段 $\mathcal{H}^s$ 所需的左/右声学上下文（定义 2）。缺少边缘编码会促使注意力解码在计算交叉注意力时，寻找其他信息来源来理解声学编码的顺序，而不再依赖分段边界。计算注意力编码器-解码器损失及交叉注意力时，我们只使用扩展后与 $\mathcal{Y}^s$ 相关的有效子集 $\mathcal{H}_E^s$，并分别舍弃为 $\mathcal{X}_L^s$ 和 $\mathcal{X}_R^s$ 计算的左侧 $\mathcal{H}_L^s$ 和右侧 $\mathcal{H}_R^s$编码（见图 1 中的阴影区域）。

分段拼接（SC）： LF声学录音使我们能够拼接任意数量的连续片段及其对应的连续声学特征 $\mathcal{X}$（包括拼接片段附近的所有非语音音频）。我们发现，这能通过以下方式提升注意力解码：i）让注意力接触更多样的片段时长；ii）让注意力解码接触更多LFE编码。 [14]近期也采用了类似方法。分段拼接与声学上下文不同：后者会移除边界条件，而前者会改变训练片段长度，且两者可以联合应用。

语义分割（SS）： 秉持着利用连接时序分类来辅助注意力解码 [8,25], 并将分段与解码相结合的思路 [26] , 我们明确使用 CTC 头来对分割标记 segE 建模。这些标记用于指示语义连贯句子的边界，而不是依赖纯音频语音活动检测。CTC 发出该标记后，会用它来确定需要进行第二遍注意力解码的片段 $\mathcal{H}^s$。请注意，注意力解码能够解码 LFE $\mathcal{H}^s \in \mathcal{H}$ 的任意子集，这是本研究的主要贡献之一；若不具备这一能力，就需要重新计算编码 $\mathcal{H}^s$ 以引入边界提示（原因详见第 3.1 节和第 4.1 节）。我们使用 任意文本分段 模型 [27]为训练转录文本添加分割标记。

4. 实验

本研究中的所有模型均基于 CTC-AED 架构 [28]。该架构部署灵活，既可用于流式处理，也可用于离线处理或两遍系统。我们训练了两种模型尺寸，分别称为 *Ours.base* 和 *Ours.small*，参数量分别约为 9000 万和 2.4 亿。两种模型共用相同的 $R = 6$倍时间卷积下采样前端，以及由 3 个单向 Transformer块组成、总参数量为 1800 万的注意力解码器 $\mathcal{D}_\theta$ 架构。声学编码器 $\mathcal{E}_\theta$ 基于因果Conformer块 [15][1] 构建，参数量分别为 7000 万和 2.19 亿。两者的维度均为 512，均包含 8 个注意力头，但块数和前馈网络（FF）维度有所不同。*Ours.base* 包含 12 个块，前馈单元为 2048 个；*Ours.small* 包含 28 个块，前馈维度为 3072。Whisper [6] 等架构会在编码器和解码器之间平均分配参数，而我们扩展了编码器，并将解码器固定为 1800 万参数。这样既能实现高效的解码器，以支持端侧自回归解码，也能打造更强大的编码器，更好地利用神经网络加速器。注意力解码和 CTC 头对相同的 6081 个字节对编码子词单元 [29]上的概率分布进行建模。

我们使用 Adam优化器，通过混合公开可用的数据来训练模型，共进行 300k 次更新，每次更新使用 6144 个句子，平均长度为 33 秒。训练数据由可通过 LDC[2], SpeechOcean、LibriHeavy [30]获取的英语语料库，以及一个从公开可访问来源收集的大规模长格式对话音频数据集组成，该数据集称为 SpeechCrawl [31]。SpeechCrawl 使用由开源模型 [6, 16] 和内部模型组成的集成模型生成伪标签，随后通过 ROVER委员会投票 [32]进行筛选。并非所有数据集都包含长篇连续声学录音，因此我们对 LibriHeavy、SpeechCrawl 和 SpeechOcean 应用长篇转换（第 3.2 节）。我们在公开基准测试上报告结果，其中包括用于长格式解码的 *Tedlium3* [33] 和 *Earnings21* [34]，以及用于短格式解码的 *Com- monVoice* [35], *Librispeech* [36]。

4.1. 结果

表1系统地展示了第2节中提出的各项改动对解决长篇声学编码（LFE）上的注意力解码问题所产生的影响。我们报告了三种解码模式的结果：注意力重打分 [9] (重打分)、注意力 [1] (注意力)以及 CTC与注意力联合解码 [37] (*CTC*-注意力)。这三种模式都以不同方式使用注意力解码，因此在解码 LF 数据时表现各异。除非另有说明，SFE 和 LFE 的分段方式相同，且对于模型 0–4，均由独立的语音活动检测系统生成。结果显示，SFE 与 LFE 上的性能差异显著，凸显了将注意力解码用于转写长格式音频片段所面临的挑战。

基线性能（模型 0）表1的第一部分报告了基线 AED 模型的词错误率（WER）。我们可以看到：i）对于 SFE，从注意力重打分改为自回归解码后，准确率显著提升（从 5.8% 提高到 4.9%）；ii）对于注意力重打分，从 SFE 改为 LFE 也有益处（从 5.8% 提高到 5.4%，这归功于质量更高的第一遍 CTC假设）。然而，对于采用注意力解码器或 CTC注意力解码器的 LFE，性能则显著下降。这表明，在分段数据上训练的原始注意力解码与 LFE 表示不兼容（定义1）。进一步调查发现，模型难以输出序列结束标记，导致反复对相同的块进行重新解码，从而产生过多的插入（仅受解码器设置的最大 token 数量限制）。

表1：所提出的改动对短篇和长篇声学编码（*AE*）上的注意力解码的影响。在 *Ours.base* 模型的 *TED-LIUM3* 长格式测试集上报告词错误率（%）。

Mdl	AE	修改				词错误率（%）		
		SC	AC	PE	SS	*AR*	*AD*	*CAT*
0	SFE	✗	✗	✗	✗	5.8	5.1	4.9
	LFE					5.4	295	14.5
1	SFE	✓	✗	✗	✗	5.5	4.9	4.8
	LFE					5.1	295	16.6
2	SFE	✓	✓	✗	✗	5.6	4.9	4.8
	LFE					5.1	40.3	8.4
3	SFE	✓	✗	✓	✗	5.6	5.0	4.8
	LFE					5.1	145	5.5
4	SFE	✓	✓	✓	✗	5.5	4.9	4.7
	LFE					5.1	5.0	4.5
5	LFE	✓	✓	✓	✓	4.7	4.8	4.3

SC - 分段拼接，AC - 声学上下文，PE - 位置编码，SS - 语义分割，AR - 注意力重打分，CAT - CTC-注意力联合解码

分段拼接（SC，模型 1）分段拼接是我们引入的第一项修改，旨在通过使用更长的多片段语句（本研究中最长可达 2 分 30 秒）进行训练，提升 LF 数据上的注意力解码性能。可以看到，对于 SFE，该方法在所有解码模式下都能降低整体词错误率，但未能解决 LFE 片段上的关键注意力解码失败问题，即那些未保留边界提示的片段（定义 2）。

声学上下文的影响（AC，模型 2）在训练期间引入声学上下文，可在保持分段性能的同时缓解基线模型在 LFE 上的困难，将词错误率从 295% 降至 40.3%；不过，与 SFE 相比，性能差距仍然很大。扩展的声学上下文促使模型减少对边界提示的依赖，更多地利用声学嵌入中已有的相对位置信息，该信息通过现有输入 RoPE [23] 位置编码提供。对于 *Ours.base* 模型，$L = 12$, $N = 16$, $M = 16$ 以及 $R = 6$，因此 $C_L^{\max} = 1152$ 和 $C_R^{\max} = 96$ 声学帧。为获得 LFE，我们分别追加 11.52 秒和 0.96 秒的左侧与右侧上下文（另见第 2.2 节）。

位置编码效应（PE，模型 3）在 $\mathcal{H}^s$（模型 3）中添加辅助绝对位置编码，带来了适度的改进：与模型 1 相比，长篇注意力解码的词错误率大约减半，降至 145%。尽管其效果不如声学上下文，位置编码仍应能提供必要的时间结构信息，帮助注意力解码理解长音频中的序列顺序。有趣的是，仅仅将这些编码附加到输入中，并不足以让模型充分利用它们，这可能是因为分段训练数据中的边界提示更容易成为学习时间顺序的捷径。不过请注意，这些辅助位置码确实优于输入中已有的旋转位置编码（参见模型 1）。

尽管模型 3 的仅注意力解码结果表明，声学上下文扩展（模型 2）带来的词错误率降幅大于分段位置编码（模型 3），但深入分析后会发现，这两种方法之间的差异更为微妙。

[1]我们使用层归一化而非批归一化，并采用旋转位置编码。 [2] https://catalog.ldc.upenn.edu/。我们使用了 Fisher、Callhome 和 ISCI

表 2：最佳系统的词错误率结果涵盖分段基准测试和长篇形式的公开基准测试。

模型	解码器	块大小（秒）	分段				长篇	
			LibriSpeech 清洁集 –	*LibriSpeech 其他集* –	*Common Voice*	*Tedlium3*	*Tedlium3*	*Earnings21*
Whisper base.en [6]	AD	30	4.1	9.6	17.5	4.6	4.6	12.5
Ours.base	AR	0.96	2.3	5.9	16.3	4.8	4.7	13.4
Ours.base	AD	0.96	2.2	5.5	14.1	4.8	4.8	13.3
Ours.base	CAT	0.96	2.1	5.4	14.4	4.4	4.3	12.2
Ours.base	CAT	3.84	1.9	4.9	13.7	4.3	4.3	12.1
Whisper small.en [6]	AD	30	3.2	6.7	12.6	4.3	4.6	10.8
Ours.small	AR	0.96	2.0	4.8	13.4	4.6	4.5	12.5
Ours.small	AD	0.96	2.0	4.6	12.1	4.4	4.4	12.1
Ours.small	CAT	0.96	1.8	4.4	12.4	4.1	4.0	11.4
Ours.small	CAT	3.84	1.7	3.9	11.4	3.9	3.9	11.4

AD - 注意力解码，AR - 注意力重打分，CAT - CTC-注意力联合解码

位置编码（模型 3）使注意力机制能够解码出更连贯、更长的句子，但模型难以输出 EOS标记。相比之下，声学扩展（模型 2）更频繁地生成 EOS标记，但代价是删除错误增加。

这些发现表明，如果能更好地指导模型何时停止解码，模型 3便可能取得更优的整体性能。我们使用联合*CTC*-注意力解码 [37], 对这一假设进行了测试，结果见表1最后一列。在 CTC 的引导下，采用辅助位置编码的注意力解码模型（模型 3），其准确率高于训练时使用扩展声学上下文的模型（模型 2）：词错误率分别为 5.5% 和 8.4%。

上下文与位置编码组合（模型4）然而，将声学上下文与位置编码结合（模型4）带来了显著提升，使 SFE 和 LFE 编码在采用注意力解码处理长音频时表现持平。这一结果表明，空间信息（声学上下文）和时间信息（位置编码）互为补充，二者对于长音频注意力解码都不可或缺；同时，在训练过程中让模型接触更多不带边界提示的数据，会促使其更多依赖辅助位置编码。这一结果还表明，添加明确的位置编码信息 $\mathcal{H}^s$ 至关重要（比较表1中模型4和模型2的注意力解码结果）。

包含所有修改的完整系统（模型5）模型5表明，使用第一遍连接时序分类语义分段标记来触发注意力解码，始终优于基于VAD的方法。这一提升在预料之中，因为注意力解码器主要基于正确分段的语句进行训练，而VAD可能会引入非语义边界，造成推理不匹配。在我们的长格式实验中，这些VAD不一致通常会导致删除错误增加。

CTC-注意力混合解码器在所有解码模式下都取得了最佳的整体性能。值得注意的是，模型5可在LFE空间中自然运行；使用VAD分段时，模型4则是与之对应的SFE基线。在Earnings-21长格式测试集上也观察到了类似的性能趋势，但此处未报告相关结果。

4.2. 最终结果

表 2 显示，我们提出的系统在分段和长格式评估数据集上均取得了出色的表现。值得注意的是，为实现有效的长篇处理而进行的调整，并未降低系统在标准分段音频上的表现，这表明我们的架构修改能够适应不同类型的分段方式。

就参数量而言，我们的模型与规模相近的Whisper变体性能相当，同时延迟更低。在长篇数据集上，我们的模型达到Whisper base.en的性能，在分段任务上则表现更佳；而我们的小型模型在大多数基准测试中始终优于Whisper small.en。请注意，*Librispeechlbs clean*、*LibriSpeech* 其他集 和*CommonVoice* 对于Whisper而言属于零样本，但对我们的系统并非如此，这或许可以解释为什么我们的模型在这三个基准测试上的表现更好。

使用 CAT 解码还带来了额外的实际优势。这种混合方法通常能取得最佳性能（例如，在 *LibriSpeech* 清洁集 上的词错误率为 1.8%，而纯注意力的词错误率为 2.0%），同时还能实现流式首遍推理、提供用于触发注意力解码的语义分段标签，并在具有挑战性的声学条件下提供更稳健的解码；在这些条件下，基于注意力的解码容易产生幻觉。由于我们的模型在编码器级别采用可变掩码进行训练，因此可以将模型的块大小设置得更大。表 2 报告了块大小为 3.96 秒时最佳 CAT 解码的结果。总体趋势符合预期：词错误率进一步降低，但延迟也有所增加。

5. 结论

我们提出了一系列改进方法，用于训练注意力解码器模型，使其能够处理长篇声学编码表示。通过声学上下文扩展、辅助位置编码，以及长格式分段采样与拼接，基于注意力解码器的系统在长篇声学编码表示上的转录质量可媲美对分段编码的解码，从而拓展了注意力解码器的应用方式。与需要处理固定大小输入窗口或依赖基于原始信号的语音活动检测决策的模型相比，采用首遍语义分割标签的混合式连接时序分类-注意力系统，为使用注意力解码两遍架构提供了一种低延迟方案。虽然我们已使注意力解码能够无缝处理长篇声学编码表示，但尚未探究通过不同片段之间的上下文连续性来扩展 AED 的其他方法 [6, 7]。我们将这一工作留待未来开展。所述进展有助于改进多项面向用户的功能，例如实时翻译、来电筛选、通话保持助手、实时语音留言、可视语音留言与通话转录、FaceTime实时字幕与系统级实时字幕、备忘录中的音频转录、语音备忘录与日志，以及 Final Cut Pro 中的转录为字幕。

6. 致谢

感谢 Nikos Flemotomos、Erik McDermott 和 Barry Theobald 为本研究提出的宝贵建议。感谢 Shaoen Qin 对 SpeechCrawl 进行大规模伪标注。

7. 参考文献

[1] J. Chorowski、D. Bahdanau、D. Serdyuk、K. Cho 等人，“用于语音识别的注意力模型”，载于 NIPS 论文集，NIPS’15 系列，2015 年，第 577–585 页。[2] W. Chan、N. Jaitly、Q. Le 和 O. Vinyals，“听、关注并拼写：用于大词汇量对话语音识别的神经网络”，载于 ICASSP 论文集，2016 年，第 4960–4964 页。[3] C.-C. Chiu、W. Han、Y. Zhang、R. Pang 等人，“长篇语音识别端到端模型比较”，载于 ASRU 论文集。IEEE，2019 年，第 889–896 页。[4] A. Narayanan、T. N. Sainath、R. Pang、J. Yu 等人，“用于统一流式与非流式自动语音识别的级联编码器”，ICASSP 论文集，第 5629–5633 页，2020 年。[5] J. D. Fox、D. Raj、N. Delworth、Q. McNamara 等人，“长篇语音识别的更新语料库与基准测试”，载于 ICASSP 论文集。IEEE，2024 年，第 13 246–13 250 页。[6] A. Radford、J. W. Kim、T. Xu、G. Brockman 等人，“通过大规模弱监督实现稳健语音识别”，arXiv 预印本 arXiv:2212.04356，2022 年。[7] T. Hori、N. Moritz、C. Hori 和 J. L. Roux，“基于 Transformer 的长上下文端到端语音识别”，载于 Interspeech 2020，2020 年，第 5011–5015 页。[8] ——，“先进的长上下文端到端语音识别：使用上下文扩展型 Transformer”，载于 I nterspeech 论文集。ISCA，2021 年，第 2097–2101 页。[9] Z. Yao、D. Wu、X. Wang、B. Zhang 等人，“WeNet：面向生产的流式与非流式端到端语音识别工具包”，载于 Interspeech 论文集。IEEE，2021 年，第 4054–4058 页。[10] T. Wang、Y. Fujita、X. Chang 和 S. Watanabe，“基于分块非自回归模型的流式端到端自动语音识别”，载于 Interspeech，2021 年，第 2266–2270 页。[11] T. Parcollet、R. van Dalen、S. Zhang 和 S. Bhattacharya，“SummaryMixing：用于语音识别与理解的线性复杂度自注意力替代方案”，2023 年，arXiv:2307.07421。[12] K. Kim、F. Wu、Y. Peng、J. Pan 等人，“E-branchformer：采用增强合并的 Branchformer，用于语音识别”，载于 SLT 论文集。IEEE，2023 年，第 84–91 页。[13] X. Zhang、Q. Zhang、H. Liu、T. Xiao 等人，“语音中的 Mamba：探索自注意力的替代方案”，IEEE TASLP，2025 年。[14] C. Carvalho、J. Tian、W. Chen、Y. Peng 等人，“探索线性变体 Transformer 与 k-NN 记忆推理在长篇自动语音识别中的应用”，载于 Interspeech，2025 年，第 3583–3587 页。[15] A. Gulati、J. Qin、C. Chiu、N. Parmar 等人，“Conformer：用于语音识别的卷积增强型 Transformer”，载于 Interspeech 论文集。ISCA，2020 年，第 5036–5040 页。[16] D. Rekesh、N. R. Koluguri、S. Kriman、S. Majumdar 等人，“采用线性可扩展注意力的快速 Conformer，实现高效语音识别”，载于 ASRU 论文集。IEEE，2023 年，第 1–8 页。[17] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit 等人，“Attention Is All You Need”，载于 NeurIPS 论文集，2017 年，第 6000–6010 页。[18] Q. Zhang、H. Lu、H. Sak、A. Tripathi 等人，“Transformer 转导器：一种采用 Transformer 编码器和 RNN-T 损失的可流式语音识别模型”，载于 ICASSP 论文集，2020 年，第 7829–7833 页。[19] T. Likhomanenko、Q. Xu、G. Synnaeve、R. Collobert 等人，“CAPE：使用连续增强位置嵌入对相对位置进行编码”，载于 NeurIPS 论文集，第 34 卷，2021 年，第 16 079–16 092 页。[20] J. Pan、T. Lei、K. Kim、K. J. Han 等人，“SRU++：开创结合注意力的快速循环，用于语音识别”，载于 ICASSP 论文集，2022 年，第 7872–7876 页。[21] X. Chen、Y. Wu、Z. Wang、S. Liu 等人，“在大规模数据集上开发用于语音识别的实时流式 Transformer 转导器”，载于 ICASSP 论文集，2021 年。[22] P. Swietojanski、S. Braun、D. Can、T. F. da Silva 等人，“用于可配置 Transformer 转导器语音识别的可变注意力掩码”，载于 ICASSP 论文集。IEEE，2023 年。[23] J. Su、M. Ahmed、Y. Lu、S. Pan 等人，“Roformer：采用旋转位置嵌入的增强型 Transformer”，Neurocomput.，第 568 卷，C 期，2024 年 2 月。[24] W. Chen、T. Kano、A. Ogawa、M. Delcroix 等人，“长时间训练、长时间测试：在语音处理中利用完整文档上下文”，载于 ICASSP 论文集，2024 年，第 13 066–13 070 页。[25] N. Moritz、T. Hori 和 J. Le Roux，“用于端到端语音识别的触发式注意力”，载于 ICASSP 论文集，2019 年，第 5666–5670 页。[26] W. R. Huang、S. yiin Chang、D. Rybach、R. Prabhavalkar 等人，“E2E Segmenter：面向长篇自动语音识别的联合分段与解码”，载于 Interspeech，2022 年。 [在线]。网址：https://api.semanticscholar.org/CorpusID:248366641[27]M. Frohmann、I. Sterner、I. Vuli´c、B. Minixhofer 等人，“分割任意文本：一种稳健、高效且适应性强的通用句子分割方法”，载于 EMNLP 论文集，2024 年，第 11 908–11 941 页。 [在线]。网址：https://aclanthology.org/ 2024.emnlp-main.665[28] S. Watanabe、T. Hori、S. Kim、J. R. Hershey 等人，“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。[29] R. Sennrich、B. Haddow 和 A. Birch，“使用子词单元对罕见词进行神经机器翻译”，载于 ACL 论文集，2016 年，第 1715–1725 页。[30] W. Kang、X. Yang、Z. Yao、F. Kuang 等人，“Libriheavy：包含标点与大小写信息及上下文的 50,000 小时自动语音识别语料库”，2023 年。[31] Z. Gu、T. Likhomanenko 和 N. Jaitly，“Omni-router：在用于语音识别的稀疏专家混合中共享路由决策”，arXiv 预印本 arXiv:2507.05724，2025 年。[32] J. Fiscus，“用于降低词错误率的后处理系统：识别器输出投票错误减少（ROVER）”，载于 ASRU 论文集，1997 年，第 347–354 页。[33] F. Hernandez、V. Nguyen、S. Ghannay、N. Tomashenko 等人，“TED-LIUM 3：数据量翻倍及语料库重新划分，用于说话人自适应实验”，载于 SPECOM 论文集，2018 年，第 198–208 页。[34] M. Del Rio、N. Delworth、R. Westerman、M. Huang 等人，“Earnings-21：真实场景下自动语音识别的实用基准测试”，载于 Interspeech 2021。ISCA，2021 年 8 月，第 3465–3469 页。 [在线]。网址：http://dx.doi.org/10.21437/Interspeech.2021-1915[35] R. Ardila、M. Branson、K. Davis、M. Kohler 等人，“Common Voice：一个大规模多语言语音语料库”，载于 LREC 论文集，2020 年 5 月，第 4218–4222 页。[36] 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 ICASSP 论文集，2015 年，第 5206–5210 页。[37] T. Hori、S. Watanabe 和 J. Hershey，“用于端到端语音识别的联合连接时序分类/注意力解码”，载于 ACL（第1卷：长篇论文）论文集。ACL，2017 年，第 518–529 页。