基于连接时序分类解码器的实时语音翻译方法

张绍磊[1),3)] 冯 洋[1),2),3)]

[1)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[2)](中国科学院智能算法安全重点实验室 北京 100190)

[3)](中国科学院大学 北京 100049)

摘 要 实时场景中的跨语言沟通是全球化进程中的重要场景。实时语音翻译旨在通过计算机在说话者讲话的同时输出目标语言的翻译文本,在诸多实时场景中具有广泛的应用前景。当前的离线模型尽管拥有大规模参数,但其架构仍无法直接处理实时跨语言沟通场景。在此背景下,实时语音翻译对于实时性的独有要求使得其在研究和应用上具备特定的必要性。与离线语音翻译相比,实时语音翻译更具挑战性,因为其需要额外制定读/写策略以控制模型在合适的时机开始翻译,从而在低延时下获得高质量翻译。理想情况下,实时语音翻译模型应在接收到相关语音后立即生成对应的目标文本,以确保高翻译质量和低延时。因此,建模源语音和目标文本之间的对齐是指导读/写策略的关键。基于此,本文提出了一种基于连接时序分类解码器的实时语音翻译方法。该方法通过连接时序分类技术插入空白标记和重复标记,实现语音和文本不等长序列间的对齐,并根据此对齐制定读/写策略来控制模型在接收到对应的语音之后开始翻译。在训练中引入连接时序分类损失能有效地将对齐学习与目标文本生成整合在统一的框架中,从而找到最佳的读/写策略。本文在两个实时语音翻译基准上对提出的方法进行了全面评估,结果表明提出的方法在实时语音翻译性能上超过了现有最佳方法。进一步的分析实验展示了该方法的有效性和优越性。

关键词 实时翻译;语音翻译;机器翻译;连接时序分类;非自回归生成;对齐

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.01100

CTC-Based Decoder-only Simultaneous Speech Translation

ZHANG Shao-Lei[1),3)] FENG Yang[1),2),3)]

[1)](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

[2)](Key Laboratory of AI Safety, Chinese Academy of Sciences, Beijing 100190)

[3)](University of Chinese Academy of Sciences, Beijing 100049)

Abstract Cross-lingual communication in real-time scenarios is a key aspect of the globalization process. Simultaneous speech translation (SimulST) aims to output the target-language translation concurrently with the speaker's speech, offering promising applications in various real-time scenarios. Although current offline models have large-scale parameters, their architescture still cannot directly address real-time cross-lingual communication needs. In this background, the unique requirements of real-time performance make SimulST particularly necessary for both research and practical applications. Unlike offline speech translation, SimulST is more challenging due to the necessity of a READ/WRITE policy that controls the model to start translating at appropriate moments, thereby achieving high-quality translation with low latency. Ideally, a SimulST model should generate the corresponding target text immediately upon receiving the a-

收稿日期:2024-08-08;在线发布日期:2025-02-26。本课题得到国家自然科学基金项目(No. 62376260)资助。**张绍磊**,博士研究生,主要研究领域为自然语言处理、实时翻译和大语言模型。E-mail:zhangshaolei20z@ict.ac.cn。**冯洋(通信作者)**,博士,研究员,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为自然语言处理和大语言模型。E-mail:fengyang@ict.ac.cn。

ligned speech inputs, ensuring both high translation quality and low latency. Therefore, modeling the alignment between the source speech and the target text is essential for guiding the READ/WRITE policy. In this paper, we introduce a decoder-only SimulST model (DeST) based on Connectionist Temporal Classification (CTC) alignments. DeST learns the alignments between source speech and target text using the CTC loss, and then determines the READ/WRITE actions based on this alignment. CTC loss can effectively integrate learning alignment and generating target text in a unified framework during training, thereby finding the optimal READ/WRITE policy. The experimental results on two speech translation benchmarks show that the proposed method outperforms previous strong simultaneous speech translation baselines. Further analyses demonstrate the effectiveness and superiority of the proposed method.

Keywords　simultaneous translation; speech translation; machine translation; connectionist temporal classification; non-autoregressive generation; alignment

1　引　言

实时场景中的跨语言沟通是全球化进程中的重要场景。实时语音翻译(Simultaneous Speech Translation,SimulST)旨在将源语言语音即时转化为目标语言文本[1-7]。该技术在国际会议、实时直播、跨国旅游等诸多实时场景中得到了广泛应用,为人们提供了低延时的跨语言交流服务,成为机器翻译领域的研究热点[8]。随着大模型技术的发展,大模型显著提升了许多自然语言处理任务的表现。但在低延时的实时跨语言沟通场景下,超大参数量的大模型由于其较慢的推理速度仍未能提供理想的解决方案,使得实时语音翻译的研究在低延时需求场景下显得尤为重要。与此同时,实时语音翻译研究的进展,将为未来如何构建能够实现高效实时沟通的大型语言模型提供重要的思路和启示。

实时语音翻译面临着巨大的挑战,因为它要求在低延时的情况下生成高质量的翻译。与离线语音翻译需要等待完整的源语音输入不同,实时语音翻译需要一个读/写策略来控制模型在接收流式输入时决定继续等待源语言语音输入(读操作)还是开始翻译目标语言文本(写操作),如图 1 所示。理想情况下,实时语音翻译模型应在接收到相关的源语言语音后再开始生成对应的目标语言文本,从而确保较高的翻译质量。因此,目标语言文本和源语言语音之间的对齐可以用于指导实时语音翻译的读/写策略。

现有的实时语音翻译读/写策略主要分为两类:固定策略和自适应策略。固定策略根据预先制定的

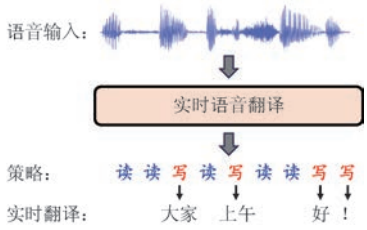


图 1　实时语音翻译示意图

规则执行读/写操作,例如 Wait-k 策略在先等待固定时间的语音输入后,每接收到 280 毫秒语音生成一个目标单词。此类方法完全忽略了目标语言文本和源语言语音的相关性,可能破坏源语言语音的完整性,并强制模型在未收到相关源语言语音时输出目标语言文本,从而导致翻译质量下降。自适应策略则根据当前输入的语音动态决策读/写操作,从而实现更好的实时语音翻译性能。然而,现有的自适应策略通常只关注源语言语音和源语言文本之间的对齐来找到合适的切分位置,例如依赖外部语音分段模型或启发式的语音边界检测器来切分源语言语音,却忽略了源语言语音和目标语言文本之间的对齐关系。这使得其在一些语法结构差异较大的语音间进行实时语音翻译时,难以找到最佳的翻译时机。

为此,本文旨在探究一种基于源语言语音和目标语言文本之间对齐的实时语音翻译方法。考虑到实时语音翻译的延时受到模型开始翻译时机和推理速度的双重影响,本文提出了一种基于连接时序分类解码器(CTC-based Decoder-only Simultaneous Speech Translation,DeST)的实时语音翻译方法,其通过连接时序分类技术将对齐和生成整合在统一的框架下。DeST 采用连接时序分类解码器进行实时语音翻译,其整体为非自回归(Non-autoregres-

sive)解码器架构,并利用连接时序分类(Connectionist Temporal Classification,CTC)技术实现不等长的语音和文本序列之间的对齐。具体而言,DeST 通过连接时序分类解码器直接将源语言转化为目标语言文本,并在目标文本词表中插入特殊空白标记ϵ,当模型生成ϵ时,说明没有目标词对齐到当前的语音输入,DeST 继续等待新的语音输入(即执行读操作);反之,当模型生成目标词时,则说明该目标词对齐到当前输入上,模型可以开始翻译生成目标文本(即执行写操作)。在训练过程中,带有ϵ和重复标记的目标序列通过 CTC 损失函数与标准译文进行优化,从而联合学习对齐和目标文本的生成。如此一来,DeST 能够基于源语言语音和目标语言文本之间的对齐获得准确的读/写策略,另外,连接时序分类解码器的非自回归结构也将带来较快的推理速度,从而在低延时下获得高质量翻译。

本文在公开的语音翻译数据集上评估了 DeST 的性能,包括 MuST-C 英语到德语语音翻译数据集和 MuST-C 英语到西班牙语语音翻译数据集。实验结果表明,与现有的固定策略和自适应策略相比,DeST 方法在相同延时下取得了更好的翻译质量。进一步的分析实验验证了提出的 DeST 在实时语音翻译任务中的有效性和优势。DeST 为在大模型尚未能解决的实时跨语言场景中的实时语言翻译提供了解决方案,同时为未来的实时交互模型的研究和应用提供了重要的参考价值。

整体上,本文的贡献主要体现在以下三方面:

(1)本文提出了一种基于连接时序分类的读/写策略,实验结果表明该策略能够在更加准确的时机开始翻译,从而取得更好的实时翻译质量。

(2)本文提出了一种基于非自回归解码器的语音翻译模型,实验表明该架构在离线语音翻译性能上与编码器—解码器架构相媲美,并且具有更快的推理速度。

(3)本文通过全面的实验和分析验证了提出方法中各个模块的有效性和优越性。最后,本文给出了提出方法面临的局限性以及潜在的未来研究。

2 相关工作

本节将介绍本文所涉及的相关工作,包括语音翻译和实时语音翻译。

2.1 语音翻译

语音翻译(Speech Translation)是指将源语言的语音输入转换成目标语言文本输出的技术。语音翻译任务的数据形式通常表示为语音—源语言文本—目标语言文本三元组 $\{(S,X,Y)\}$,其中 $S=(s_1,\cdots,s_{|S|})$ 是源语言语音,$X=(x_1,\cdots,x_{|X|})$ 是语音对应的源语言转录文本,$Y=(y_1,\cdots,y_{|Y|})$ 是目标语言翻译文本。语音翻译的方法可以分为级联方法和端到端方法。

(1)级联方法。早期的语音翻译系统通常采用级联方法[9-12],即先将源语音转录成文本(语音识别)[13],然后将文本翻译成目标语言(机器翻译)[14]。这种方法的优点在于可以利用已有的语音识别和机器翻译技术,达到一定的翻译效果。但其缺点在于容易在两个阶段之间积累误差,导致翻译质量下降[15]。此外,级联方法在处理不同语言的语音翻译时,可能面临语言特定问题,如语言模型和声学模型不匹配等,这进一步影响了翻译的准确性。

(2)端到端方法。近年来,端到端(end-to-end)的语音翻译模型逐渐兴起[16-17]。这种模型直接从源语言的语音信号生成目标语言文本,不经过中间的文本表示,因而可以减少误差积累[18],提高翻译质量。典型的端到端语音翻译模型基于序列到序列架构,这些模型通常包含一个编码器(encoder)和一个解码器(decoder)。编码器将源语音信号编码成隐层表示,解码器则根据这些隐层表示生成目标语言的文本输出。当前语音翻译研究中,最广泛采用的模型架构为 Transformer 架构,其由 Vaswani 等人[19]于 2017 年提出并在序列建模任务中展现了出色的性能。整体的端到端语音到文本翻译任务可以表示为 $p(Y|S)$。端到端方法由于其直接性和连贯性,能够更好地捕捉语音信号中的上下文信息,从而提升翻译的自然度和准确性。此外,近年来一些语音翻译工作探索了使用非自回归架构(non-autoregressive architectures)[20-21]或者仅使用解码器(decoder-only architectures)[22-23]来完成语音到文本的翻译,取得了极具潜力的效果。

2.2 实时语音翻译

实时语音翻译(Simultaneous Speech Translation)是指将流式输入的源语言语音实时转换成目标语言文本并同步输出的技术。该技术在国际会议、跨国商务交流以及多语言直播等场景中具有广泛的应用前景。

端到端实时语音翻译模型除了执行语音翻译,还需要一个策略来控制模型执行读/写操作,其中读操作控制模型继续等待实时语音输入,而写操作控

制模型生成一个目标词。形式化的,用 $g(i)$ 表示模型生成第 i 个目标词的时机。因此,生成 y_i 的概率为 $p(y_i \mid S_{\leqslant g(i)}, Y_{<i})$,其中 $S_{\leqslant g(i)}$ 为当前接收到的源语言语音,$Y_{<i}$ 为之前生成的目标词。实时语音翻译模型需要制定策略来决定 $g(i)$,从而在合适的时机产生高质量翻译结果。根据其读/写策略的不同,端到端方法可以分为固定策略和自适应策略。

固定策略　固定策略按照预设规则执行读/写操作,能够在保证一定翻译质量的同时简化决策过程。Ma 等人[24]提出了一种固定预决策方法,将语音分割为 280 毫秒的等长片段,然后使用 Wait-k 策略(Wait-k Policy)[4]或者单调多头注意力(Monotonic Multi-head Attention,简称 MMA)[25]来决策读/写操作。固定策略的优点在于其实施简便且易于调试,然而,由于其缺乏对输入动态变化的适应能力,可能在处理复杂语音信号和不同类型语言输入时表现出一定的局限性。

自适应策略　自适应策略根据输入的动态特征灵活调整读/写操作,以实现更高效和准确的翻译效果。Ren 等人[26]提出了 SimulSpeech,通过检测源语言语音中的单词数量对语音进行分割,然后执行 Wait-k 策略。Chen 等人[27]通过源语言语音的语音识别结果来决策读/写操作,进一步提升了实时翻译的准确性。Zeng 等人[28]提出了 RealTrans,通过检测源语音包含的单词数来缩短语音片段长度并进行实时翻译,从而减少延迟。Dong 等人[29]提出了 MoSST,在声学信息累计超过一定阈值后进行翻译,提高了系统对长句子的处理能力。Zhang 等人[30]提出了 ITST,通过判断接收到的语音信息是否足够翻译来动态决策读/写操作。Zhang 等人[31]提出了 MU-ST,基于语义单元对源语言语音进行分割并进行实时语音翻译。

3　方　法

决策翻译时机是实时语音翻译任务的核心挑战。为了确保较高的翻译质量,实时语音翻译模型应在接收到足够的源语言语音后,再开始生成对应的目标语言文本。因此,源语言语音和目标语言文本之间的对齐关系可以有效地指导最佳翻译时机。为此,本文提出了一种基于连接时序分类解码器的实时语音翻译方法(CTC-based Decoder-only Simultaneous Speech Translation,DeST),通过连接时序分类技术来建模语音和文本间的对齐关系,并依据此对齐关系来决策何时开始翻译。本节将详细介绍 DeST 的模型架构、训练和推理过程。

3.1　模型架构

图 2 展示了本文提出的连接时序分类解码器架构,该架构整体采用非自回归结构,由声学特征提取器和解码器[14]组成。声学特征提取器从原始语音输入中提取语音特征,解码器将语音特征以非自回归的方式映射为目标文本,并通过连接时序分类(CTC)技术来学习目标语言文本和源语言语音之间的对齐,从而决策何时开始翻译。各模块具体介绍如下。

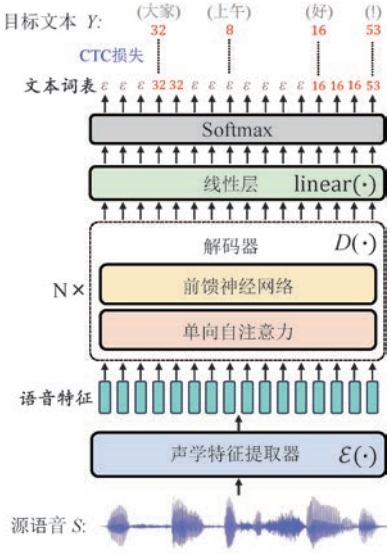


图 2　基于连接时序分类解码器的实时语音翻译示意图

声学特征提取器　声学特征提取器用于从输入的源语言语音 S 中提取语音特征。本研究采用预训练的 Wav2Vec 2.0 模型[32],该模型通过堆叠的卷积层和 Transformer 层将原始语音信号转换为高维的语音特征表示。具体地,声学特征提取过程可以表示为公式(1)所示。

$$H = \mathcal{E}(S) \tag{1}$$

其中,$\mathcal{E}(\cdot)$ 表示声学特征提取器,本文采用预训练的 Wav2Vec 2.0 模型。S 表示原始语音输入,$H = (h_1, \cdots, h_{|H|})$ 表示从原始语音输入提取到的语音特征序列。在训练过程中,声学特征提取器的参数处于冻结状态。需要注意的是,声学特征提取涉及下采样操作和卷积操作,因此其提取过程会压缩原始语音序列的长度。对于基于 Wav2Vec 2.0 模型的声学特征提取器,每个语音特征对应于 20 毫秒的原始语音输入,即每接收到 20 毫秒语音即可产生一个新的语音特征。解码器在提取语音特征之后,解码器(Decoder)以非自回归的方式对语音特征进行

上下文建模并解码出目标文本。解码器的输入为语音特征 H，输出是目标语言文本。解码器包含 N 个堆叠的标准 Transformer 解码器层[14]，每个 Transformer 层由单向自注意力(self-attention)和前馈神经网络(feed-forward network)两个子层构成。单向自注意力通过计算语音特征之间的注意力来捕捉上下文信息。为了实现对实时语音的建模，自注意力机制采用单向方式，即每个位置的语音特征只能关注该设置之前的语音特征。形式化地，语音特征 h_i 只能通过注意力关注到那些 $j \leqslant i$ 的语音特征 h_j，以此来满足对流式语音建模的实时性要求。前馈神经网络由两个线性变换层和一个激活函数组成，能够对隐藏状态进行进一步的特征变换，从而增强模型的表示能力。形式化地，解码器对语音特征的建模过程可以表示为公式(2)所示。

$$D = \mathcal{D}(H) \tag{2}$$

其中，$\mathcal{D}(\cdot)$ 表示解码器，D 表示经过解码器处理后的隐藏状态。

读/写策略　设计一个有效的读/写策略是实时语音翻译模型在低延时下实现高质量翻译的关键。具体而言，一个理想的读/写策略应该控制模型在接收相关的语音特征之后输出目标文本，确保翻译质量的同时避免等待过多语音输入导致的高延时。所以，如果能将每个目标语言文本对齐到源语言语音上，目标词所对齐位置即为最佳的开始翻译的时机。

然而，目标文本序列和源语音序列长度往往存在明显差异，为此，本研究采用连接时序分类(CTC)技术来建模源语音和目标文本之间的对齐。CTC 技术是实现不等长序列间对齐的常用技术，其允许模型在输出序列中包含重复标记和空白标记，而生成目标词的位置即为对齐位置。通过 CTC 技术，源语音和目标文本之间的对齐和目标文本的生成被整合在统一的框架下完成。具体而言，为了通过 CTC 技术建模不等长的源语音和目标文本序列之间的对齐关系，本文在原始的文本词表 V 中引入空白标记 ϵ，得到带有空白标记的新词表 V'，如公式(3)所示。

$$V' = V \cup \{\epsilon\} \tag{3}$$

基于隐藏状态 D，模型通过线性层将高维的隐藏状态 D 转换为词表大小 $|V'|$ 的向量，并使用 Softmax 层转化为词表 V' 上的概率分布，用于预测每个位置上可能的词汇。

在推理过程中，将模型在第 i 个语音特征 h_i 处生成的目标词记为 y_i。如果 y_i 为空白标记 ϵ，则说明没有目标词和语音特征 h_i 对齐，模型继续等待之后的语音输入(即读操作)；反之，如果模型在 h_i 处生成的文本标记 y_i 不是空白标记 ϵ 或者重复标记，则说明 y_i 对齐到语音特征 h_i，模型可以输出 y_i(即写操作)。形式化地，DeST 在解码出 y_i 后执行的读/写操作可以表示为公式(4)所示。

$$\text{Action} = \begin{cases} \text{READ}, & \text{if } y_i = \epsilon \text{ or } y_i = y_{i-1} \\ \text{WRITE}, & \text{otherwise} \end{cases} \tag{4}$$

其中，y_{i-1} 为在之前第 $i-1$ 位置输出的目标词。如果执行读操作(READ)，模型将继续等待新的语音输入；如果执行写操作(WRITE)，模型将输出 y_i。考虑到可能存在源语言语音结束后目标文本可能尚未生成完成的情况，本研究在源语言语音结束后持续填充静音语音(即全 0 的语音特征向量)。这种方式模拟了人类译员在语音结束后能持续进行翻译的模式，支持模型在源语言语音停止后继续生成额外的目标标记。通常情况下，目标语言文本的生成位置与语音特征是相对齐的，因此仅有少数几个单词可能在源语言语音结束后仍未生成。在实践中，本研究在语音输入结束后持续补充静音直到模型生成结束符〈EOS〉(即结束生成文本)之后停止，用于处理语音结束后仍未生成的单词部分。最终，整个实时翻译过程将在模型生成了结束符〈EOS〉之后停止。

3.2 训　练

源语言语音特征序列 H 和目标文本序列 Y 是两个不等长的序列，这使得按照对应位置应用交叉熵损失进行训练变得困难。为了克服这一挑战，本文提出了一种基于连接时序分类(Connectionist Temporal Classification，CTC)损失的模型进行训练。CTC 损失常用于非自回归结构中，以在不等长的序列间建立有效的监督机制。通过这种方式，模型能够同时学习源语音和目标文本之间的对齐关系，以及目标文本的生成过程。

CTC 引入一个扩展的输出序列 Z，其长度为 $|H|$，并允许在目标序列中插入特殊的空白标记 ϵ 和重复标记。在测试时，通过对输出序列 Z 依次进行合并相邻的重复标记并去掉空白标记的 ϵ 操作，以得到目标序列 Y。CTC 的训练目标是最大化所有可能扩展序列 Z 的概率之和，如公式(5)所示。

$$p(Y \mid H) = \sum_{Z \in \pi^{-1}(Y)} p(Z \mid H) \tag{5}$$

其中，π 是将扩展序列 Z 映射到目标序列 Y 的压缩操作，合并相邻的重复标记并去掉空白标记 ϵ。这使得模型可以灵活地处理不同长度的输入输出序

列,从而更好地学习它们之间的对齐关系。$p(Z \mid H)$ 通过 3.1 小节中介绍的解码器来建模,其中语音特征序列 H 通过多层解码器层$\mathcal{D}(\cdot)$进行处理,最后通过线性层和 Softmax 层生成输出序列 Z 。

(1)CTC 损失。在训练过程中,输入的语音特征序列 H 和输出的目标语言文本序列 Y 之间的 CTC 损失函数 $\mathcal{L}_{\mathrm{CTC}}^{S\to Y}$ 可以表示为公式(6)和(7)。

$$\mathcal{L}_{\mathrm{CTC}}^{S\to Y} = -\log p(Y \mid H) \tag{6}$$

$$\text{其中 } H = \mathcal{E}(S) \tag{7}$$

通过最小化 CTC 损失函数,模型学习到输入序列 H 和目标序列 Y 之间的对齐关系以及目标序列 Y 的生成过程。

(2)多任务学习。除了学习从源语言语音转换为目标语言文本的翻译任务外,本文还引入了语音识别作为辅助任务进一步提升模型的翻译性能。在近来的研究中,引入语音识别等辅助任务已经被证明能够有效提升语音翻译性能。因此,本文采用多任务学习方法,在语音翻译($S\to Y$)和语音识别($S\to X$)任务上进行联合训练。为了实现多任务学习,本文采用独立的线性层和 Softmax 层来分别处理语音识别任务和语音翻译任务,同时共享声学特征提取器和 N 层解码器层。通过多任务学习,提出的方法既能通过参数共享实现多种任务的相互促进,又能够通过独立的线性层和 Softmax 层实现不同语言的解码。对于语音识别任务而言,输入同样为语音特征序列 H ,而输出为源语言的转录序列 X 。同样地,语音识别任务也通过 CTC 损失来训练,输入的语音特征序列 H 和输出的源语言文本序列 X 之间的 CTC 损失函数 $\mathcal{L}_{\mathrm{CTC}}^{S\to X}$ 可以表示为公式(8)和(9)。

$$\mathcal{L}_{\mathrm{CTC}}^{S\to X} = -\log p(X \mid H) \tag{8}$$

$$\text{其中 } H = \mathcal{E}(S) \tag{9}$$

最终,模型整体的损失函数$\mathcal{L}$为两部分 CTC 损失之和,如公式(10)所示。

$$\mathcal{L} = \mathcal{L}_{\mathrm{CTC}}^{S\to Y} + \mathcal{L}_{\mathrm{CTC}}^{S\to X} \tag{10}$$

3.3 推　理

在推理时,为了控制实时语音翻译系统的整体延时,本文引入了一个超参数——首字滞后 T。该参数用于控制模型在等待时长为 T 的语音输入之后再开始决策。更大的首字滞后使模型在更高延时下完成实时语音翻译;反之,更小的首字延时使模型以更低延时完成实时语音翻译。通过引入首字滞后,用户可以在测试过程中手动调整期望的延时水平,以适配不同延时需求的场景。

DeST 的推理算法如算法 1 所示。实时语音输入记作 S ,模型$\mathcal{M}$首先会等待时长为 T 的语音输入以确保初始语音段的完整性。这一阶段,模型仅积累语音数据,不进行任何翻译操作。此后,模型$\mathcal{M}$基于当前接收到的语音 S 生成目标词 y 。在这一过程中,模型对每一帧语音输入进行处理,并结合之前的输入进行解码,生成目标词 y 。如果生成的目标词 y 为空白标记或与之前生成过的标记重复,则模型认为当前语音帧的信息不足以生成新的翻译词,因此继续等待 20 毫秒语音输入(即 $S.READ(20\text{ms})$),然后进行下一次判断。反之,如果生成的目标词 y 不是空白标记且与之前的标记不重复,则说明目标词 y 与当前接收到的语音帧相对齐,模型输出解码出的目标词 y(即 $Y.WRITE(y)$)。然后,模型读入新的语音继续下一次判断,此时如果源语言语音输入结束,则模型读入静音。模型重复此过程直到生成结束符$\langle EOS\rangle$。

算法 1. 基于连接时序分类解码器的实时语音翻译推理算法

输入:模型$\mathcal{M}$,首字滞后 T ,实时语音输入 S
输出:目标文本 Y
初始化:$Y = [\langle BOS\rangle]$,记录重复标记 $y_{pre} = \langle BOS\rangle$

```
IF |S| < T AND not S.finished() THEN
    //当前语音长度未达到 T ,读入语音
    S.READ(20ms) ;
WHILE y ≠ ⟨EOS⟩ DO
    //解码目标标记
    y ← M.forward(S) ;
    IF ( y = ϵ OR y = y_pre ) AND not S.finished()
    THEN
        //生成空白标记或者重复标记,读入语音
        S.READ(20ms) ;
    ELSE
        //输出翻译文本
        Y.WRITE(y) ;
        //生成之后,读入新语音
        //若此时语音输入结束,读入静音
        S.READ(20ms) ;
    y ← y_pre ;
RETURN Y
```

通过这种方式,模型能够在源语言语音和目标语言文本之间建立有效的对齐,从而在那些对齐的位置生成目标词,实现低延时的高质量实时语音翻译。

4 实　验

本节对提出的方法进行了全方位的评估,将依

次介绍数据集、评价指标、基线系统和实验结果。

4.1 数据集

本文在两个实时语音翻译基准上进行了实验，包括 MuST-C① 英语到德语翻译数据集(234K 条语音－文本对)和 MuST-C 英语到西班牙语数据集[33](270K 条语音－文本对)。MuST-C 数据集是一个大规模的多语言语音翻译数据集，广泛应用于语音翻译研究。本文使用了 dev 集合作为验证集，其中英语到德语翻译的验证集包含 1423 条语音－文本对，英语到西班牙语翻译的验证集包含 1316 条语音－文本对；使用 tst-COMMON 集合作为测试集，其中英语到德语翻译的测试集包含 2641 条语音－文本对，英语到德语翻译的测试集包含 2502 条语音－文本对。

对于语音数据，本文采用了原始的 16－bit 16kHz 单声道音频波形，以确保语音信号的高质量和一致性。对于文本数据，本文使用了 SentencePiece 工具为源语言文本和目标语言文本分别生成大小为 6000 的词表。

4.2 评价指标

实时语音翻译的性能从翻译质量和延时两方面进行评估，以全面衡量模型的实际应用效果。

(1)翻译质量。本研究采用机器翻译任务的标准指标 BLEU 值[34]来评估实时语音翻译的翻译质量。BLEU 值衡量了翻译结果和标准译文之间的相似性，是一种广泛使用的自动化评估指标，能够有效反映翻译系统的性能。更进一步，为了评估翻译的语义准确性，本文使用被广泛使用的 COMET 值来评估生成的翻译和标准翻译之间的语义准确性。

(2)延时。本研究采用平均滞后[4](Average Lagging，AL)来评估实时语音翻译的延时。AL 衡量了目标语言文本输出滞后于源语言语音输入的平均时间(毫秒)。形式化地，记录模型在时刻$\mathcal{T}(y_i)$生成了目标词 y_i，则 AL 的计算如公式(11)和(12)所示。

$$\mathrm{AL}=\frac{1}{\tau}\sum_{i=1}^{\tau}\mathcal{T}(y_i)-\frac{i-1}{|Y|/T_S} \tag{11}$$

$$\tau=\mathrm{argmin}_i(\mathcal{T}(y_i)=T_S) \tag{12}$$

其中，T_S 表示语音输入的总时长，τ 表示源语言语音结束时生成的目标单词数。平均滞后 AL 值通过计算目标词生成的时间，量化了翻译过程中产生的延时。

本文应用开源的 SimulEval 工具②[35]来模拟语音实时输入的过程。该工具可以自动化模拟说话人说话并在过程中将流式语音实时发送给模型，并接受模型返回的翻译结果。最终，该工具可以自动化地计算翻译质量和延时的评估指标。

4.3 基线系统

为了全面评估提出方法的有效性，本文在以下系统上进行了实验：

(1)离线翻译(Offline)：离线语音翻译系统在接收完整的语音输入后再进行翻译，滞后时间即为完整语音输入的时长。离线翻译系统采用标准的 Transformer 模型，即编码器－解码器架构，其包括 6 层编码器层和 6 层自回归的解码器层。

(2)Wait-k：Wait-k 策略[4]是实时语音翻译中最广泛应用的策略。Wait-k 策略以固定时长(280 毫秒)对语音进行分段[24]，并设置每 280 毫秒的语音对应一个词。基于此，Wait-k 策略控制模型首先滞后 k 个语音片段(即 $k\times 280$ 毫秒)，然后每接收 280 毫秒语音就生成 1 个目标词，直到翻译结束。

(3)Wait-k-Stride-n：Wait-k-Stride-n 策略[28]是 Wait-k 策略的变体。Wait-k-Stride-n 策略控制模型首先滞后 k 个语音片段，然后每 $n\times 280$ 毫秒生成 n 个目标词。参考原文中的最佳设置，本文选择 $n=2$，即每接收 560 毫秒语音就生成 2 个目标词。

(4)MMA③：单调多头注意力机制(Monotonic Multihead Attention，MMA)[25]其将语音分割为等长的片段(120 毫秒、200 毫秒和 280 毫秒[24])，然后通过引入一个 0/1 伯努利变量来进行读/写决策。在训练中，伯努利变量以期望的形式和翻译模型通过注意力机制一起联合训练。

(5)SimulSpeech：基于单词检测器的实时语音翻译方法[26]，其通过识别源语言语音中所包含的源语言单词数来决策是否开始翻译。在训练中，SimulSpeech 引入注意力分布上的知识蒸馏技术提升翻译性能。

(6)SH：ASR 辅助的实时语音翻译方法(Shortest Hypothesis，SH)[27]其对源语言语音进行语音识别(ASR)并采用识别结果中的最短候选来指示语音中的源语言单词数，然后根据此单词数来执行 Wait-k 策略。

(7)RealTrans：卷积加权 Transformer[28]，其通过 CTC 损失来检测源语言语音所对齐的源语言单词数并结合 Wait-k-Stride-n 策略进行解码。

① https://ict.fbk.eu/must-c。

② https://github.com/facebookresearch/SimulEval。

③ https://github.com/pytorch/fairseq/examples/simultaneous_translation。

(8) MoSST①:单调分段实时语音翻译方法(Monotonic-segmented Streaming Speech Translation, MoSST)[29],其采用整合一发射方法(integrate-and-fire[36])根据累积的声学信息进行语音分段。然后基于语音分段数来执行 Wait-k 策略。

(9)ITST②:基于信息运输的实时语音翻译方法(Information-transport-based Simultaneous Translation, ITST)[30],通过量化从源语言到目标语言的传输信息,并根据接收到的累积信息决定是否进行翻译。

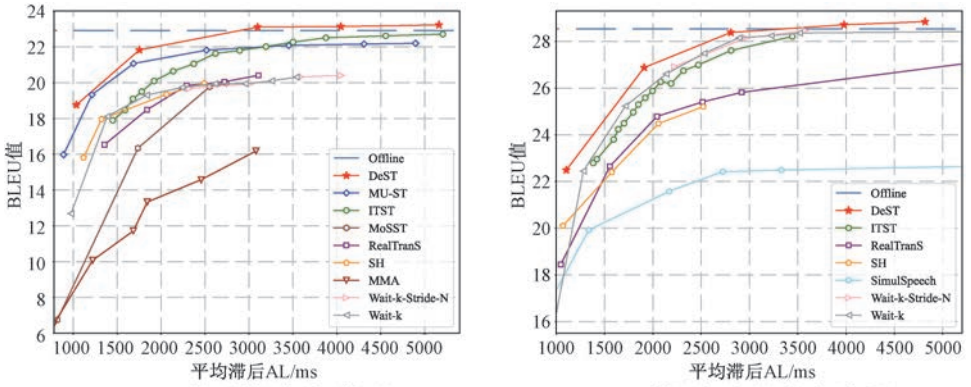
(10)MU-ST:基于语义单元的分段方法(Simultaneous Translation based on Meaning Unit, MU-ST)[31],其通过构建的数据训练外部语音分段模型,用于判断当前接收到的语音是否是完整的语义单元。MU-ST 并使用该语音分段模型来决定离线翻译模型何时进行翻译。

(11)DeST:本文提出的基于连接时序分类解码器的实时语音翻译方法。DeST 通过结合连接时序分类(CTC)和传统的翻译模型,实现了对实时语音翻译过程的精确控制和高效处理。

所有基线均基于 Fairseq 库[37]来实现。本文使用预训练的 Wav2Vec 2.0③[32]作为声学特征提取器,并采用标准的 Transformer decoder[14]作为解码器。对于解码器,语音特征维度为 768 维,多头注意力机制的注意力头数为 8,前馈神经网络中间维度为 2048。为了和标准的 6 层编码器—6 层解码器架构保持参数统一,DeST 的解码器包含 12 层 Transformer 解码器层。整个训练采用 adam 优化器,学习率为 0.0001,热身步数(warmup step)为 4000 步,dropout 设置为 0.1。本文在 4 块 NVIDIA 3090 GPU 上进行训练直到模型在验证集上的表现收敛,并最终采用验证集表现最佳的模型进行实验。

4.4　实验结果

本文在 MuST-C 英语到德语翻译数据集和 MuST-C 英语到西班牙语翻译数据集上进行了实验,以评估 DeST 的实时翻译性能。为了全面地测试 DeST 的效果,本研究将首字滞后 T 调整为不同的值,以获得不同延时下的翻译质量。然后将延时(AL,毫秒)作为横坐标,翻译质量(统计准确率 BLEU 值和语义准确性 COMET 值)作为纵坐标,绘制翻译质量—延时曲线,结果如图 3 和图 4 所示。



(a) MuST-C英语到德语实时语音翻译　　(b) MuST-C英语到西班牙语实时语音翻译

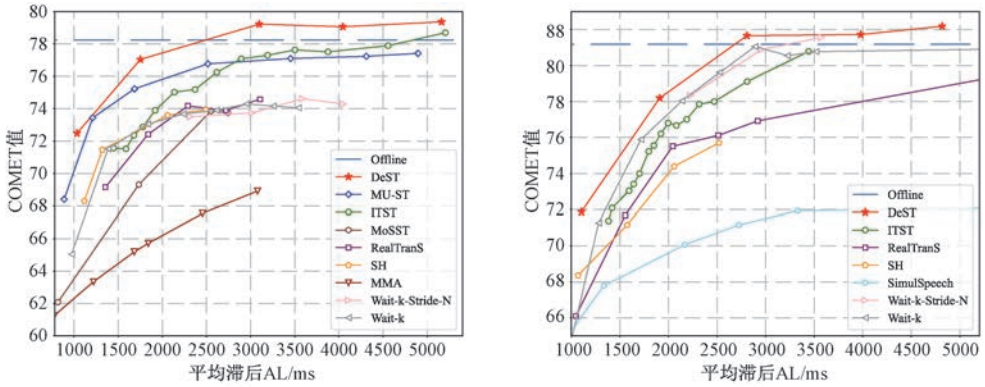
图 3　实时语音翻译性能(BLEU 值)

实验结果表明,DeST 在所有延时下均优于现有方法,并在相同延时下取得了更高的实时语音翻译质量。对于固定策略,Wait-k、Wait-k-Stride-n 等方法只能根据预先制定的规则执行读/写操作,例如每 280 毫秒翻译一个词,这完全忽略了源语言语音和目标语言文本之间的相关性。由于语音信号的多样性,每个单词对应的语音长度并不固定,每 280 毫秒翻译一个词容易迫使模型在未接收到足够语音信息的情况下翻译目标词,导致翻译质量下降。相比之下,提出的方法 DeST 能够根据源语言语音和目标语言文本之间的对齐关系,动态决策执行读操作或写操作,从而在各个延时情况下均表现出显著的优势。

对于自适应策略,之前的 RealTranS、MoSST、SH、MU-ST 等方法通常只关注源语言语音和源语言文本之间的对齐(即从源语音中识别出的源语言词汇),并利用包含的源语言单词数量来控制模型执行读/写操作。这类方法相较于固定策略有所改进,但基于源语言单词数量的策略与翻译目标单词之间

① https://github.com/dqqcasia/mosst。
② https://github.com/ictnlp/ITST。
③ https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt。

仍存在不匹配的问题。尤其在语序和结构差异较大的语言对间进行实时翻译时(例如从主谓宾结构的英语翻译到主宾谓结构的德语时),源语言单词数量难以准确反映能生成的目标语言单词数量。根据语音中包含的源语言单词数量制定读/写策略难以获得特别准确的翻译时机。本文提出的 DeST 方法通过直接建模源语言语音和目标语言文本之间的对齐关系,并利用此对齐信息来决策读/写操作。这种方式能直接控制模型在目标语言文本的对齐位置上开始翻译,从而在实时语音翻译质量上表现更佳。



(a) MuST-C英语到德语实时语音翻译 (b) MuST-C英语到西班牙语实时语音翻译

图 4 实时语音翻译性能(COMET 值)

除了在实时语音翻译质量方面的优势,DeST 的另一个显著特点是其更快的推理速度,这对于实时语言翻译至关重要。表 1 展示了 DeST 与基于标准 Transformer 架构的实时语音翻译模型之间的推理速度对比。以往的实时语音翻译方法通常采用编码器解码器架构的 Transformer 模型,每秒能够生成约 10 个目标标记。其中 MMA 和 MU-ST 等方法由于引入了额外模块,推理速度相比离线模型降低。而本文提出的 DeST 方法则采用非自回归解码器架构,每秒能够生成约 73 个目标标记,推理速度相比于离线模型提高了约 7 倍。标准的 Transformer 模型采用编码器—解码器架构,其中解码器通过自回归方式逐词解码目标文本,因此需要多次循环解码器。与此不同,DeST 方法采用非自回归解码器架构,对于语音输入仅需进行一次推理,即可生成多个目标文本,因此在推理速度上具有显著优势。

多语言设置。DeST 结构支持多语言设置,因此,本节将评估其在多语言设置下的表现。具体而言,在多语言设置中,训练使用了 MuST-C 英语到德语翻译和 MuST-C 英语到西班牙语翻译的混合语料,且各语言的词表共享,统一训练 DeST 模型。在推理阶段,单一 DeST 模型能够同时完成 MuST-C 英语到德语和英语到西班牙语的翻译任务。图 5 的实验结果表明,语言间参数共享进一步提升了 DeST 模型在不同语言上的性能,进一步验证了 DeST 结构在多语言设置下的有效性。

表 1 模型推理速度对比

方法	模型架构	英语到德语翻译		英语到西班牙语翻译	
		推理速度(标记/秒)	加速比	推理速度(标记/秒)	加速比
Offline	编码器—解码器	10.22	1.00	11.03	1.00
Wait-k	编码器—解码器	10.24	1.00	11.04	1.00
Wait-k—Stride-n	编码器—解码器	10.24	1.00	11.04	1.00
MMA	编码器—解码器	3.24	0.32	4.12	0.37
SimulSpeech	编码器—解码器	9.12	0.89	10.20	0.92
SH	编码器—解码器	10.19	1.00	10.97	0.99
RealTrans	编码器—解码器	8.99	0.88	9.38	0.85
MoSST	编码器—解码器	9.75	0.95	10.29	0.93
ITST	编码器—解码器	9.92	0.97	10.37	0.94
MU-ST	编码器—解码器	9.03	0.88	9.85	0.89
DeST	非自回归解码器	73.23	7.17	81.15	7.36

注:其中加速比为相对于离线模型(offline)的加速倍数。

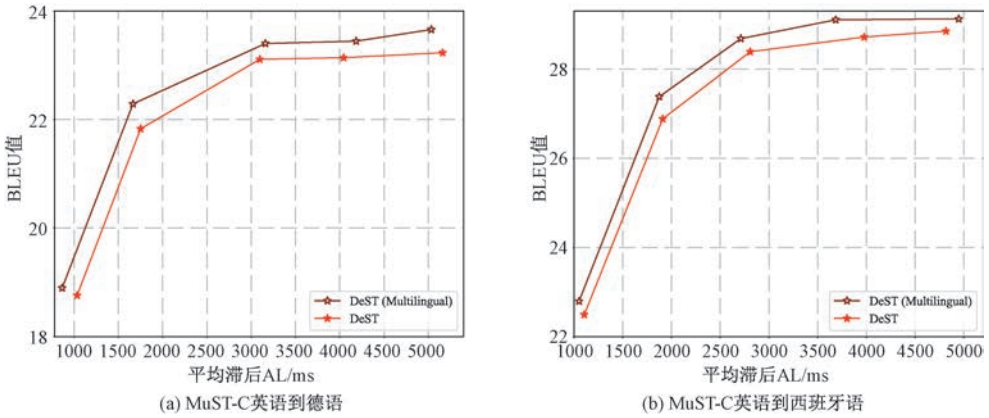


图 5　DeST 在多语言设置下的性能提升

综上所述,本节通过实验验证了 DeST 在实时语音翻译任务中具有显著的性能提升和推理速度优势,为实际应用提供了有力支持。

5　分　析

本节通过充分的实验分析,深入探讨了提出方法的有效性和优越性。为了公平对比,本节中所有分析实验均在单语种设置下进行。

5.1　与基于 Transformer 的翻译模型间的比较

为了探究提出的方法相比基于 Transformer 的语音翻译模型的优势,本节在离线场景下评估了 DeST 和基于 Transformer 的语音翻译模型在离线语音翻译上的性能。本文提出的 DeST 模型不仅能够实现实时语音翻译,还具备离线语音翻译的能力。具体而言,通过在测试时将首字延时设置为无穷大(即 $T=\infty$,在完整语音输入结束后再开始翻译),DeST 模型可以完成离线语音翻译。

(1)性能。为了验证 DeST 的离线语音翻译性能,本节在表 2 中报告了 DeST 模型在离线语音翻译任务中的表现。结果表明,相较于传统基于编码器—解码器架构的离线语音翻译模型,DeST 模型在离线语音翻译性能方面取得了一定提升,平均提升约 0.4 BLEU 值。这一性能提升的主要原因在于基于解码器的架构(decoder-only)在相同参数量的情况下具备更多的层数,从而具有更强的建模能力。这表明基于解码器的架构对于语音翻译任务具有一定潜力,为未来的语音翻译研究提供了新的思路和方向。

(2)效率。表 1 展示了报告了 DeST 和基于 Transformer 的语音翻译模型(Offline)之间的推理速度。DeST 相比于基于 Transformer 的语音翻译模型平均具有 7 倍左右的加速比,具备明显的效率用优势。更快的推理速度为 DeST 在实时场景中的应用提供了优势。整体上,DeST 在性能和效率上相比基于 Transformer 的语音翻译模型均具有一定优势。

表 2　离线语音翻译性能对比(BLEU 值)

方法	参数量	英语到德语	英语到西班牙语
Offline	190M	22.92	28.54
DeST	178M	23.34	28.9
Δ	—	+0.42	+0.36

5.2　读/写策略的有效性

本文提出的方法 DeST 在实时语音翻译任务中取得了显著的性能提升。为了深入探讨实时语音翻译性能的提升是否源于提出的基于对齐的读/写策略的精确性,本节在 MuST-C 英语到德语翻译数据集上对 DeST 的读/写策略进行消融实验。实验通过将 DeST 中基于目标语言文本和源语言语音对齐的读/写策略替换为 Wait-k 策略(记为 DeST(Wait-k Policy))、MU-ST 策略(记为 DeST(MU-ST Policy))和 ITST 策略(记为 DeST(ITST Policy)),来验证本文提出的基于对齐的读/写策略的有效性。具体而言,DeST 保持模型结构不变,但读/写策略不再依赖生成空白标记或重复标记进行决策,而是采用滞后固定时长的 Wait-k 策略、基于外部切分的 MU-ST 策略,以及基于运输信息量的 ITST 策略。

图 6 展示了对 DeST 读/写策略进行消融实验的结果。实验表明,在相同模型架构下,当 DeST 的读/写策略被替换为固定策略(Wait-k 策略)时,模型的实时翻译性能显著下降。特别是在低延时(滞后 1000 毫秒)的情况下,性能下降尤为明显,BLEU 值下降约 2.5。与此同时,当 DeST 的读/写策略被替换为之前的自适应策略(如 MU-ST 策略和 ITST

策略)时,模型的实时翻译性能也出现了不同程度的下降。这些实验结果进一步验证了 DeST 中基于对齐的读/写策略在实时翻译中的关键作用,该策略能够确保模型在适当的时机开始翻译目标词,从而有效提升整体翻译性能。

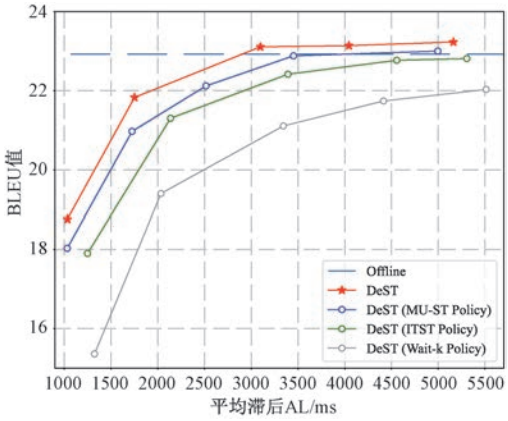


图 6 DeST 中读/写策略的消融实验

5.3 多任务学习的有效性

DeST 的训练过程采用了多任务学习方法,通过引入语音识别作为辅助任务来提升语音翻译的效果。为了验证多任务学习方法的有效性,本节在 MuST-C 英语到德语翻译数据集上进行消融实验,对比了采用多任务学习进行训练的模型和仅使用语音翻译任务(即去掉多任务学习,即为 DeST w/o multitask learning)进行训练的模型在实时语音翻译上的性能,实验结果如图 7 所示。结果展示了两种训练方法在实时语音翻译任务中的性能差异。具体来说,当去除了语音识别辅助任务后,语音翻译的性能下降了约 1.5 个 BLEU 值。这一结果表明,多任务学习方法通过共享参数能够有效地提升模型的语音翻译质量。此外,该结果也与之前许多关于离线语音翻译研究的结论相一致,进一步验证了多任务学习在语音翻译领域的潜力和优势。

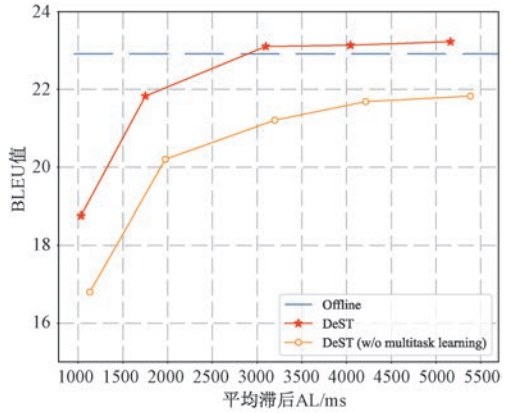


图 7 DeST 中读/写策略的消融实验

5.4 读/写策略质量评估

为了探究 DeST 的读/写策略是否能在合适的时机开始翻译,从而不破坏语音的完整性,本文借助带标注的语音分割数据集——Buckeye 语料库①[38]来评估翻译时机的适切性。Buckeye 语料库在语音中标注了每个词对应的边界。在实时语音翻译中,模型应在接收完整的语音信息后开始翻译,即在这些边界处开始翻译;反之,如果在语音片段中开始翻译,则容易破坏语音的完整性,影响翻译质量。因此,这些边界往往是理想的翻译时机,评估实时语音翻译模型在这些边界上的命中率可以反映读/写策略的质量。

值得注意的是,语音分割数据集中的边界是以语音帧为单位的(即某一帧是边界),而在这些边界附近的一些帧开始翻译也可能同样能确保翻译时的语音完整性,即不错的翻译时机。然而,为了延续之前工作中的实验设置,本研究仍然评估开始翻译时机与标注语音边界帧完全命中的准确率,这可以对模型的翻译时机进行更严格的评估。

本研究沿用了语音分割任务上的评估指标[39],包括精确度(Precision)、召回率(Recall)、F1 值、过分割率(Over-Segmentation,OS)和 R-value。精确度、召回率及其对应的 F1 值用于衡量模型的翻译时机与真实边界位置的一致性。过分割率(OS)[40]则用于评估模型生成的分割数量的准确性,其计算如公式(13)所示。

$$\mathrm{OS}=\frac{Recall}{Precision}-1 \tag{13}$$

其中,当 OS=0 时,表示分割数量完全准确;OS 值越大,说明生成的分割数量越多;OS 值越小,说明生成的分割数量越少。由于在生成较多分割时容易获得较高的召回率,但往往伴随较低的精确度,因此 R-value[41]被提出用于综合衡量召回率和过分割率。R-value 的计算如公式(14)~(16)所示。

$$\text{R-value}=1-\frac{|r_1|+|r_2|}{2} \tag{14}$$

$$\text{其中 } r_1=\sqrt{(1-Recall)^2+\mathrm{OS}^2} \tag{15}$$

$$r_2=\frac{-\mathrm{OS}+Recall-1}{\sqrt{2}} \tag{16}$$

一个更大的 R-value 表示更好的分割质量,只有在召回率达到完美(即 $Recall=1$)且过分割率为零(即 OS=0)的情况下,才能达到最佳 R-value。本

① https://buckeyecorpus.osu.edu。

研究对比了 DeST 策略和 Wait-k 策略、MUST 策略、ITST 策略的翻译时机质量,并提供了一些之前语音分段方法[39,42-44]的分段质量作为参考,结果如表 3 所示。实验结果表明 DeST 能够在更加准确的位置开始翻译。具体而言,相比于之前的实时语音翻译策略,DeST 取得了约 9% 的 F1 值提升。DeST 通过目标语言文本和源语言语音之间的对齐动态地决策读/写,能够更倾向于在语音的边界处开始翻译,从而确保开始翻译时接收到的语音信息是相对完整的,这有利于实时语音翻译性能。与直接的语音分段方法相比,DeST 的翻译时机基本达到了之前语音分段方法的分段水平,表明 DeST 能够在理想的翻译时机处开始翻译。综上所述,本小结通过系统的实验验证了 DeST 策略在实时语音翻译中的有效性。

表 3　实时语音翻译策略的翻译时机质量对比

方法		Precision (↑)	Recall (↑)	F1 (↑)	OS (0)	R-value (↑)
语音分段方法	ES K-Means[42]	30.7	18.0	22.7	−41.2	39.7
	BES GMM[43]	31.7	13.8	19.2	−56.6	37.9
	VQ-CPC DP[39]	18.2	54.1	27.3	196.4	−86.5
	VQ-VAE DP[39]	16.4	56.8	25.5	245.2	−126.5
	DSegKNN[44]	30.9	32.0	31.5	3.5	40.7
实时语音翻译策略	Wait-k Policy[4]	28.1	16.3	20.7	−42.0	38.4
	MU-ST Policy[31]	31.2	16.3	21.4	−47.8	39.1
	ITST Policy[30]	26.7	19.3	22.4	−27.7	38.6
	DeST Policy	34.1	28.4	31.0	−16.7	43.8

注:结果均为百分比。

5.5　案例分析

为了更清晰地展示实时语音翻译过程,本文在图 8 和图 9 中可视化了 DeST 在 MuST-C 英语到德语的实时翻译过程。为清晰起见,源语言语音所对应的转录文本被标注在语音信号上方,可视化结果展示了 DeST 在每一时刻接收到的语音输入对应输出的目标词。除此之外,其余时刻模型均生成空白标记ϵ或重复标记,并继续等待语音输入。

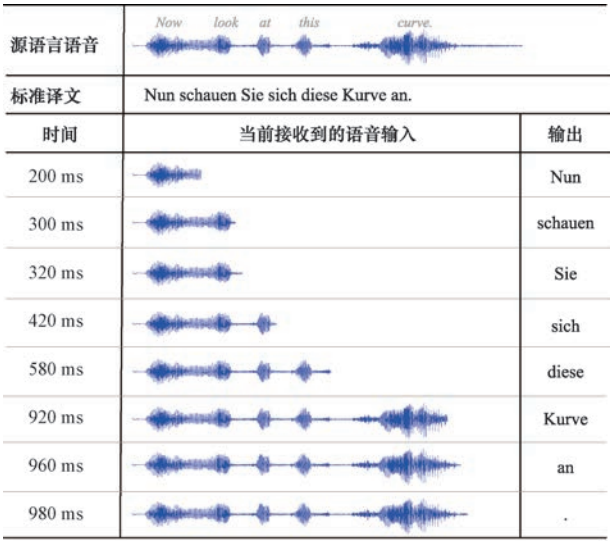


图 8　DeST 处理具有相同语序的英语到德语翻译案例时的实时翻译过程

(1)具有相同语序的翻译案例

图 8 展示了 DeST 在具有相同语序的英语到德语案例上的实时翻译过程。DeST 往往在接收到对应的源语言语音之后再开始生成目标词,例如在接收语音“Now”之后生成其对应的德语“Nun”,接收语音“look”之后生成其对应的德语“schauen”,接收语音“curve”之后生成其对应的德语“Kurve an”。这一现象体现出本文基于源语言语音和目标语言文本之间对齐所提出的读/写策略的优越性。DeST 的翻译过程依赖于源语言语音和目标语言文本之间的对齐,模型会在接收到对应的源语言语音信号后再生成目标词。

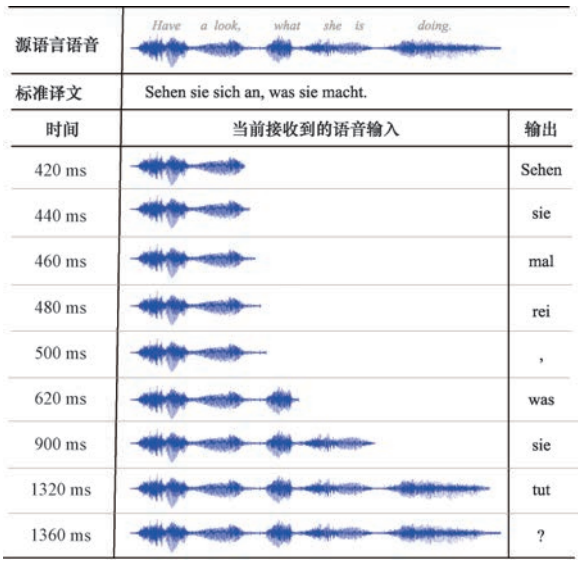


图 9　DeST 处理具有不同语序的英语到德语翻译案例时的实时翻译过程

(2)具有不同语序的翻译案例

图 9 展示了 DeST 在处理具有不同语序的英语

到德语翻译案例时的实时翻译过程。语序差异是实时翻译模型面临的一个重要挑战。例如,英语和德语在句法结构上存在显著差异,尤其是动词在句子中的位置可能不同。以图 7 中的例子为例,英语中的“look”位于第一个子句的结尾,而其对应的德语翻译“Sehen”则位于德语句子的开头。在遇到这种局部语序调换的翻译时,DeST 在前 400 毫秒内保持等待(即生成空白标记ϵ),直到接收到“look”及相对完整的源语音语义后,才开始输出“Sehen”。得益于 DeST 采用的基于对齐的策略,模型能够在等待至接收到相对完整的语义信息后再生成对齐的目标文本,从而有效应对不同语言间的语序差异。

整体上,DeST 能够在准确的位置开始翻译目标词。这种基于对齐的读/写策略促使模型在接收到对齐的源语音之后即可翻译目标词,从而确保了较高的翻译质量,同时避免了不必要的额外延时。在某些时刻,当模型尚未接收到足够的语音输入时,它会生成空白标记ϵ或重复标记。这种机制允许模型在必要时保持等待状态,避免生成错误的翻译结果。总而言之,DeST 在进行实时语音翻译时,能够动态调整生成目标词的时机,从而在低延时下取得了较高的翻译质量。

6 总结与未来研究

本文提出了一种基于连接时序分类(CTC)解码器的实时语音翻译方法。该方法通过 CTC 技术建模源语言语音与目标语言文本之间的对齐,并基于此对齐制定读/写策略,从而在低延时下实现高质量实时语音翻译。两个公开的实时语音翻译数据集上的实验结果表明,提出的方法在实时语音翻译性能上优于现有方法,并在相同延时下取得更好的翻译质量。本文的研究填补了当前实时语音翻译研究中对源语言语音与目标语言文本之间对齐建模的空白,为在当前大模型尚未能胜任的跨语言实时沟通场景提供了新的思路和技术路径,具备充分的研究和应用价值。本文方法的局限性主要体现为确保模型能够在实时场景中快速响应,当前模型在参数量上仍处于较少水平。因此,当前模型只能完成实时语言翻译这一项特定任务。为了拓展其应用范围,未来的研究将着重于如何通过增加模型的参数量,进一步提升其在实时语音交互等复杂场景中的表现能力,同时确保推理的实时性。总的来说,本文的研究成果不仅为解决跨语言实时沟通问题提供了新的技术方案,也为未来大模型的实时能力构建奠定了基础。通过本文的工作,相关技术的探索与应用将有助于推动大模型在实际生产环境中的落地应用,为日后实时交互任务的实现和优化提供了宝贵的启发和指导。

致 谢 我们衷心感谢各位专家在审稿过程中对本论文提出的宝贵意见!

参 考 文 献

[1] Guan Mo-Lin, Sun Jing, Ma Lan, et al. Principle and implementation method of computer speech real-time translation system. Examination Weekly, 2011(83): 163-164 (in Chinese)
(关墨霖, 孙晶, 马兰, 等. 计算机语音实时翻译系统原理和实现方法 . 考试周刊, 2011(83): 163-164)

[2] Grissom II A, He H, Boyd-Graber J, et al. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 1342-1352

[3] Gu J, Neubig G, Cho K, et al. Learning to translate in real-time with neural machine translation//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain, 2017: 1053-1062

[4] Ma M, Huang L, Xiong H, et al. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 3025-3036

[5] Li Tian-Yun. Analysis of machine simultaneous interpretation system under the model of interpreting work. Oriental Translation, 2018,000(6): 34-39 (in Chinese)
(李天韵. 口译工作模型下的机器同声传译系统分析. 东方翻译, 2018, 000(6): 34-39)

[6] Guo Hui-Jun. Machine simultaneous translation system based on artificial intelligence technology and speech recognition. Modern Electronic Technology, 2022, 045(009): 152-156 (in Chinese)
(郭慧骏. 基于人工智能技术和语音识别的机器同步翻译系统. 现代电子技术, 2022, 045(009): 152-156)

[7] Lu Xin-Chao. Human and machine simultaneous interpretation: Comparison and prospects of cognitive processes, abilities, and quality. Chinese Translators Journal, 2023, 44(3): 135-141 (in Chinese)
(卢信朝. 人工与机器同声传译: 认知过程, 能力, 质量对比与展望. 中国翻译, 2023, 44(3): 135-141)

[8] Li Ya-Chao, Xiong De-Yi, Zhang Min. A survey of neural ma-

chine translation. Chinese Journal of Computers, 2018, 41(12): 2734-2755 (in Chinese)
(李亚超, 熊德意, 张民. 神经机器翻译综述. 计算机学报, 2018, 41(12): 2734-2755)

[9] Stentiford F W, Steer M G. Machine translation of speech. British Telecom Technology Journal, 1988, 6(2): 116-122

[10] Wu Z, Caglayan O, Ive J, et al. Transformer-based cascaded multimodal speech translation//Proceedings of the 16th International Conference on Spoken Language Translation. Hong Kong, China, 2019: 1-8

[11] Bentivogli L, Cettolo M, Gaido M, et al. Cascade versus direct speech translation: Do the differences still make a difference? //Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 2873-2887

[12] Bahar P, Bieschke T, Schlüter R, et al. Tight integrated end-to-end training for cascaded speech translation//2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen, China, 2021: 950-957

[13] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China, 2016: 4960-4964

[14] Utskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks//Proceedings of the 2014 Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 3104-3112

[15] Ruiz N, Federico M. Assessing the impact of speech recognition errors on machine translation quality//Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track. Vancouver, Canada, 2014: 261-274

[16] Wang C, Wu Y, Liu S, et al. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 34. New York, USA, 2020: 9161-9168

[17] Wang C, Wu Y, Liu S, et al. Curriculum pre-training for end-to-end speech translation//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 3728-3738

[18] Etchegoyhen T, Arzelus H, Gete H, et al. Cascade or direct speech translation? a case study. Applied Sciences, 2022, 12(3): 1097-1108

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 2017 Advances in Neural Information Processing Systems. Long Beach, USA. 2017: 5998-6008

[20] Chuang S P, Chuang Y S, Chang C C, et al. Investigating the reordering capability in CTC-based non-autoregressive end-to-end speech translation//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online, 2021: 10681077

[21] Xu C, Liu X, Liu X, et al. CTC-based non-autoregressive speech translation//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023: 13321-13339

[22] Wu J, Gaur Y, Chen Z, et al. On decoder-only architecture for speech-to-text and large language model integration//Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2023). Taipei, China, 2023: 1-8

[23] Huang C W, Lu H, Gong H, et al. Investigating decoder-only large language models for speech-to-text translation//Proceedings of the 25th Interspeech Conference. Kos Island, Greece, 2024: 832-836

[24] Ma X, Pino J, Koehn P. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation//Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Suzhou, China, 2020:582-587

[25] Ma X, Pino J M, Cross J, et al. Monotonic multihead attention//Proceedings of the International Conference on Learning Representations. Online, 2020: 1-16

[26] Ren Y, Liu J, Tan X, et al. SimulSpeech: Endto-end simultaneous speech to text translation//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 3787-3796

[27] Chen J, Ma M, Zheng R, et al. Direct simultaneous speech-to-text translation assisted by synchronized streaming ASR//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online, 2021: 4618-4624

[28] Zeng X, Li L, Liu Q. RealTranS: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online, 2021: 2461-2474

[29] Dong Q, Zhu Y, Wang M, et al. Learning when to translate for streaming speech//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 680-694

[30] Zhang S, Feng Y. Information-transport-based policy for simultaneous translation//GOLDBERG Y, KOZAREVA Z, ZHANG Y. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 992-1013

[31] Zhang R, He Z, Wu H, et al. Learning adaptive segmentation policy for end-to-end simultaneous translation//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 7862-7874

[32] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representa-

tions//Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. Online, 2020: 12449-12460

[33] Di Gangi M A, Cattoni R, Bentivogli L, et al. MuST-C: a Multilingual Speech Translation Corpus//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 2012-2017

[34] Post M. A call for clarity in reporting BLEU scores //Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium, 2018: 186-191

[35] Ma X, Dousti M J, Wang C, et al. SIMULEVAL: An evaluation toolkit for simultaneous translation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online, 2020: 144-150

[36] Dong L, Xu B. Cif: Continuous integrate-and-fire for end-to-end speech recognition//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain, 2020: 6079-6083

[37] Ott M, Edunov S, Baevski A, et al. fairseq: A fast, extensible toolkit for sequence modeling//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Minneapolis, USA, 2019: 48-53

[38] Pitt M A, Johnson K, Hume E, et al. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication, 2005, 45(1): 89-95

[39] Kamper H, Van Niekerk B. Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks//Proceedings of the 2021 Interspeech Conference. Brno, Czech Republic, 2021: 1539-1543

[40] Petek B, Andersen O, Dalsgaard P. On the robust automatic segmentation of spontaneous speech//Proceedings of the Fourth International Conference on Spoken Language Processing. Philadelphia, USA, 1996: 913-916

[41] Räsänen O, Laine U, Altosaar T. An improved speech segmentation quality measure: the r-value//Proceedings of the 10th Annual Conference of the International Speech Communication Association. Brighton, UK, 2009: 1851-1854

[42] Kamper H, Livescu K, Goldwater S. An embedded segmental k-means model for unsupervised segmentation and clustering of speech//Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop. Okinawa, Japan, 2017: 719-726

[43] Kamper H, Jansen A, Goldwater S. A segmental framework for fully-unsupervised large-vocabulary speech recognition. Computer Speech Language, 2017, 46: 154-174

[44] Fuchs T S, Hoshen Y, Keshet J. Unsupervised word segmentation using k nearest neighbors //Proceedings of the Interspeech 2022 Conference. Incheon, Republic of Korea, 2022: 4646-4650

ZHANG Shao-Lei, Ph. D. candidate. His research interests are natural language processing, simultaneous translation and large language model.

FENG Yang, Ph. D., professor. Her research interests mainly focus on natural language processing and large language models.

Background

The task of simultaneous speech translation (SimulST) lies within the broader field of machine translation and speech recognition, aiming to convert spoken language into target-language text in real time. This field has garnered significant attention due to its critical applications in international conferences, live broadcasts, and cross-cultural communication, where low-latency, high-quality translations are essential.

Significant progress has been made internationally in addressing the challenges of SimulST, with methods primarily categorized into fixed and adaptive policies. Fixed policies operate on predefined rules to manage READ/WRITE actions, providing a straightforward approach but often struggling with diverse and complex speech inputs. In contrast, adaptive policies dynamically adjust READ/WRITE actions based on current speech inputs, yielding better performance in handling varied and fluctuating speech signals. Despite these advancements, existing methods frequently rely on external segmentation models or heuristic boundary detectors, which neglect the crucial alignment between source speech and target text.

This paper addresses these limitations by introducing a novel CTC-based Decoder-only Simultaneous Speech Translation (DeST) method. DeST integrates the alignment and generation within a unified framework, leveraging Connectionist Temporal Classification (CTC) to achieve high-quality, low-latency speech translation. Experimental results on show that the proposed method outperforms existing methods in simultaneous speech translation, achieving better

translation quality at the same latency. Further analyses verify the effectiveness of each module in the method and the superiority of the alignment-based READ/WRITE policy.

The implications of this research extend far beyond the domain of SimulST. By advancing methodologies for real-time processing and alignment, this work offers valuable insights and potential inspiration for other real-time tasks such as streaming automatic speech recognition (ASR) and real-time text-to-speech (TTS). These tasks also require efficient handling of continuous input data and real-time generation of output, making the principles developed in this research broadly applicable. The integration of CTC and decoder-only architectures could stimulate further developments in these fields, promoting innovations in real-time human-computer interaction technologies.

Moreover, enhancing simultaneous translation technologies has significant practical implications. Improved low-latency, high-quality translations can facilitate more effective communication in multilingual settings, reduce language barriers, and enhance global collaboration. This research, therefore, holds the potential to impact various domains, including international business, education, healthcare, and beyond.

Our research group has previously contributed to this field with several studies, laying a strong foundation for this work. Building on this foundation, the present study aims to push the boundaries of simultaneous speech translation, offering both practical solutions and theoretical advancements that can drive future research and applications in related areas.