基于鲁棒思维链的大语言模型语音翻译方法

房庆凯[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 语音翻译作为促进全球信息交流与消除语言障碍的关键技术，在国际会议、在线教育等多元场景中应用日益广泛。大语言模型凭借其强大的文本理解与生成能力，为语音翻译领域带来了新的突破契机。其中，思维链技术通过引导大语言模型先生成源语言转录再进行翻译，虽提升了性能，但也易导致模型过度依赖中间文本、忽略原始语音输入，从而可能放大转录缺陷，影响最终翻译的准确性与鲁棒性。为此，本文提出一种面向语音翻译的鲁棒思维链方法。该方法在训练阶段，一方面按预设概率对思维链中的部分词元进行随机掩码，以此迫使模型减少对转录文本的绝对依赖，更多地从原始语音信号及不完整的文本线索中学习推理，提升其对不完整思维链的适应与纠错能力；另一方面则引入一种正则化机制，旨在约束模型在拥有完整思维链与面对不完整思维链两种情况下对目标译文的预测分布，力求使其保持一致，从而缓解模型在拥有完整思维链时可能发生的过度依赖问题，进而全面增强翻译的准确性与鲁棒性。在 CoVoST 2 数据集六个主要翻译方向上的系统性实验评估表明，本文所提出的方法相较于不使用思维链的基线系统，平均 BLEU 得分提升高达 2.78 分；相较于标准的思维链方法，亦能取得 0.92 BLEU 分的性能增益。此外，在不同参数规模的 Qwen2.5 系列模型上的验证结果充分证明了该方法的有效性、通用性与良好的可扩展潜力。

关键词 大语言模型；语音翻译；机器翻译；思维链；鲁棒性

中图法分类号 TP18

Robust Chain of Thoughts for Speech Translation with Large Language Models

FANG Qing-Kai[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 Speech Translation (ST), as a key technology for promoting global information exchange and eliminating language barriers, sees increasingly widespread application in diverse scenarios such as international conferences and online education. Its utility is further amplified in real-time communication platforms, cross-cultural healthcare delivery, and global business negotiations, where accurate and swift translation of spoken content is crucial. Large Language Models (LLMs), with their powerful text understanding and generation capabilities, have created new opportunities for breakthroughs in the ST field. These models, trained on vast and diverse textual corpora, exhibit emergent abilities like contextual reasoning, which are highly beneficial for complex language tasks. Within this context, Chain-of-Thought (CoT) techniques, by guiding LLMs to first generate source language transcriptions before translation, have improved performance but also

本课题得到国家自然科学基金项目(No. 62376260)资助. 房庆凯，博士研究生，主要研究领域为自然语言处理、语音翻译和大语言模型.冯洋（**通信作者**），博士，研究员，博士生导师，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.

tend to cause models to over-rely on intermediate text and neglect original speech input. This reliance creates a fragile pipeline where the model's performance becomes heavily dependent on the quality of the initial transcription step. This can amplify transcription defects and consequently affect the accuracy and robustness of the final translation. Errors such as homophone confusion, out-of-vocabulary words, or speaker accent variations in the transcription phase are often directly propagated to the translation output, leading to undesirable results. To address this challenge, this paper proposes a Robust Chain-of-Thought (Robust CoT) method for speech translation. The core philosophy of this method is to train the model to treat the CoT as a helpful but fallible reasoning aid rather than an absolute ground truth. During the training phase, this method, on one hand, randomly masks parts of the tokens in the CoT sequence with a predefined probability, compelling the model to reduce its absolute dependence on transcribed text and learn to infer more from the original speech signal and incomplete textual cues, thereby enhancing its adaptability and error-correction capabilities for imperfect CoTs. This masking strategy effectively simulates various transcription error scenarios during training, fostering a more robust model. On the other hand, it introduces a regularization mechanism designed to constrain the model's predictive distributions for target translations to be consistent under conditions of both complete and incomplete CoTs. This aims to alleviate the over-reliance issue that can occur even when a complete CoT is available, thus comprehensively enhancing translation accuracy and robustness. The joint effect of these two components ensures that the model leverages the CoT when it is accurate but can seamlessly fall back to the acoustic signal when the CoT is unreliable. Systematic experimental evaluations on six major translation directions of the CoVoST 2 dataset show that the proposed method achieves a significant average BLEU score improvement of up to 2.78 points compared to a baseline system without CoT, and also yields a performance gain of 0.92 BLEU points over the standard CoT approach. Furthermore, validation results on Qwen2.5 series models of varying parameter scales fully demonstrate the effectiveness, generality, and strong scalability potential of this method. Additional analyses reveal that the model exhibits a marked improvement in handling utterances with high word error rates in the intermediate transcription.

Key words Large Language Models; Speech Translation; Machine Translation; Chain of Thoughts; Robustness

1 引言

语音翻译（Speech Translation，简称 ST）旨在将一种语言的语音转换为另一种语言的文本，从而消除不同语言人群之间的沟通障碍，对于促进文化交流与知识传播具有重要意义。近年来，随着全球化进程的加快，无论是国际会议、在线教育、跨国贸易，还是日常的跨文化沟通，对跨语言交流的需求日益增长，语音翻译技术应运而生，并在这些场景中发挥了重要作用。传统语音翻译系统多采用级联架构[1-6]，先由自动语音识别（Automatic Speech Recognition，简称 ASR）模型将语音转录为源语言文本，再通过机器翻译（Machine Translation，简称 MT）模型[7-9]将该文本翻译为目标语言。然而，级联系统不仅存在错误累积（即语音识别的错误会直接传递并影响后续的机器翻译模块）和处理延迟较高等固有缺陷，还面临着各模块单独优化导致目标

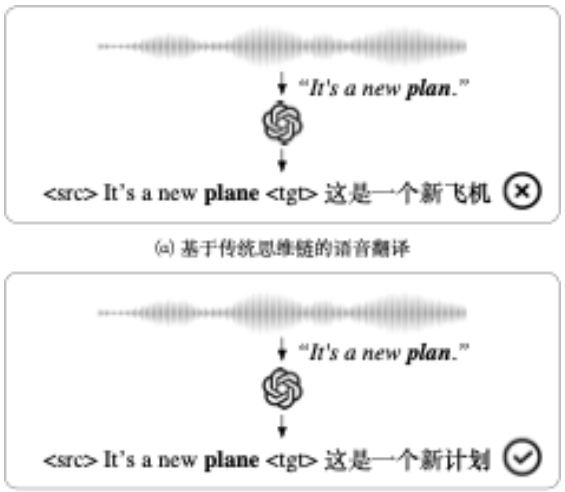


图 1 基于鲁棒思维链的语音翻译示意图

不一致、信息在模块传递间可能损失以及系统调试部署复杂等问题。针对这些挑战，端到端（End-to-End）语音翻译模型逐渐成为主流[10-20]。这类模型通常采用基于 Transformer 的编码器-解码器结构[21]，其中编码器负责对源语音信号进行表征学习，解码器则基于编码后的源语音表示直接生成目标语言文本。这种端到端联合优化的方式有效避免了级联系统中的错误叠加，简化了训练流程，并在翻译质量和响应速度上已可媲美甚至超越传统级联系统[22-25]。

随着 ChatGPT 等大语言模型（Large Language Model，简称 LLM）的出现和快速发展，采用仅解码器（Decoder-Only）结构的 LLM 逐渐成为自然语言处理领域的主流架构，展现出强大的语言理解、生成和推理能力。例如，已有研究表明，ChatGPT 在机器翻译任务中的表现已超越传统的谷歌翻译、百度翻译等专用系统[26]，其不仅在翻译的忠实度和流畅度上表现出色，产生的翻译结果也更加符合人类的语言习惯和偏好，这得益于 LLM 强大的上下文理解及自然语言生成能力。在此背景下，如何有效地将大语言模型的能力迁移并应用于构建高质量的语音翻译模型，吸引了众多研究者的关注[27-30]。为了使大语言模型支持语音这种非文本模态的输入，常见的方法是：首先利用预训练的语音编码器（如 Whisper[31]、Wav2Vec 2.0[32]等强大的声学模型）提取语音特征，然后通过线性层或 Q-Former[33]等接口模块将语音特征映射到大语言模型的输入空间，最后由大语言模型以自回归（Autoregressive）的方式生成目标语言文本。尽管此类方法在部分高资源翻译方向上的表现已优于传统语音翻译模型，但在数据资源受限的场景下，尤其对于训练语料稀疏的语种，这些模型的翻译质量仍不尽理想，存在显著的提升空间。

为进一步提升大语言模型在语音翻译任务中的表现，特别是在低资源场景下的性能，近期有研究借鉴了 LLM 在复杂文本任务上的成功经验，提出将“思维链”（Chain of Thoughts，简称 CoT）引入语音翻译流程[34-35]。其核心思想是引导大语言模型在接收语音输入后，首先自回归生成中间步骤——源语言文本（即对输入语音的转录），再进一步基于原始语音和生成的源语言文本共同生成目标语言文本。这种将语音到目标文本的直接翻译任务显式地解耦为语音到源文本和“语音+源文本”到目标文本两个阶段的方式，使得模型在翻译目标语言时可以参考更易于 LLM 理解和处理的中间文本表示，从而在多个基准上显著提升了翻译质量。然而，现有方法普遍忽视了一个潜在的关键问题：这种建模方式可能导致模型在生成最终翻译时过度依赖中间生成的源语言文本，而相应地减弱了对原始语音输入的直接感知和利用，从而可能重新引入类似于级联系统中的错误传播问题：若中间生成的源语言文本因语音识别不准或表达不自然而存在缺陷（例如，漏词、错词、重复等），这些缺陷会被直接传递到后续的翻译步骤，甚至可能被放大，最终影响翻译的准确性。如图 1 所示，这一问题在处理同音或近音词时尤为突出。例如，对于语音输入“It's a new plan”，传统的思维链方法在第一步的语音转录中，可能由于声学相似性而错误地将其识别为“It's a new plane”。在第二步的翻译阶段，模型由于过度依赖这个存在错误的中间文本，进而会将其误译为“这是一个新飞机”，如图 1(a)所示。理想的鲁棒模型则应能结合原始语音信息和上下文，即便在中间转录不完美的情况下，也能推断出正确的语义，并给出如图 1(b)所示的正确翻译“这是一个新计划”。为此，本文提出一种面向语音翻译的鲁棒思维链（Robust Chain-of-Thoughts）方法。该方法旨在充分利用思维链所带来的翻译质量提升的同时，着力缓解模型对中间源语言文本的过度依赖，并显著增强模型在面对不完美或不完整思维链时的鲁棒性。具体而言，在训练阶段，本文按一定概率对思维链中的部分词元（token）进行随机掩码，并要求模型在部分思维链信息缺失的条件下依然能够生成高质量的翻译结果，以此迫使模型减少对转录文本的绝对依赖，转而更多地从原始语音信号及不完整的文本线索中学习和推理，提升其对不完整思维链的适应与纠错能力。进一步地，本文在训练时引入一种正则化机制，其目标是约束模型在拥有完整思维链与面对不完整思维链两种条件下对目标译文的预测分布，力求使其保持一致性，从而缓解模型在拥有完整思维链时对思维链的过度依赖，从而提高翻译的准确性与鲁棒性。

为验证本文所提出方法的有效性，在广泛使用的 CoVoST 2 数据集的 6 个主要翻译方向上进行了系统性的实验评估。实验结果清晰表明，本文所提出的鲁棒思维链方法显著提升了基于大语言模型的语音翻译质量：相较于不使用思维链的直接翻译基线系统，本方法带来的平均 BLEU 得分提升高达 2.78 分；相较于标准的、未经鲁棒优化的思维链方

法，本方法亦能取得平均0.92 BLEU分的性能增益。进一步地，本文在Qwen2.5系列的不同参数规模模型（包括0.5B、1.5B、3B、7B）上均对所提方法进行了验证，实验结果充分体现了该方法在不同参数规模下的有效性、通用性与良好的可扩展性，表明其并非特定于某一模型大小，具有向更大规模模型迁移应用的潜力。

整体上，本文的贡献主要体现在以下三方面：

（1）本文首先提出了一套基于大语言模型的高效语音翻译建模方案，通过对语音特征的长度适配以及训练阶段的低秩适配，在显著降低训练成本的同时，实现了良好的基础翻译性能。

（2）本文采用思维链策略增强语音翻译，并提出了一种鲁棒思维链的训练策略，缓解模型对思维链的过度依赖，最终显著提高语音翻译的准确性与鲁棒性。

（3）本文通过全面的实验和分析，不仅验证了所提方法相较于基准模型及常规思维链方法的优越性，也证实了其在不同模型规模、不同语言和不同数据量下的通用性与有效性，并在此基础上对未来研究方向进行了展望。

2 相关工作

端到端语音翻译[36-37]能够直接将源语言语音信号转换到目标语言文本，相较于传统级联系统，在减少错误累积和降低延迟方面展现出了显著优势。然而，语音翻译领域面临的一大核心挑战，在于高质量平行数据的严重匮乏，这在很大程度上限制了相关技术的进步与应用，在方言等低资源场景下尤为突出[38]。在此背景下，借鉴并利用机器翻译领域的资源与成果，成为了应对这一挑战的重要策略。相较于语音翻译，文本翻译任务拥有更为海量且易于获取的平行语料资源，这为语音翻译模型的训练提供了宝贵的间接数据支持。为充分利用机器翻译数据，研究者们提出多种技术策略：

预训练（Pre-training）[39-42]：首先利用大规模文本翻译数据对模型的部分参数进行预训练，使其首先掌握文本翻译能力，随后在小规模语音翻译数据上进行微调以获得语音翻译能力；

多任务学习（Multi-task Learning）[17,20,43-44]：通过让模型同时训练语音翻译、机器翻译及语音识别等多个相关任务，并共享部分模型参数，以期实现跨任务的知识迁移和性能增益；

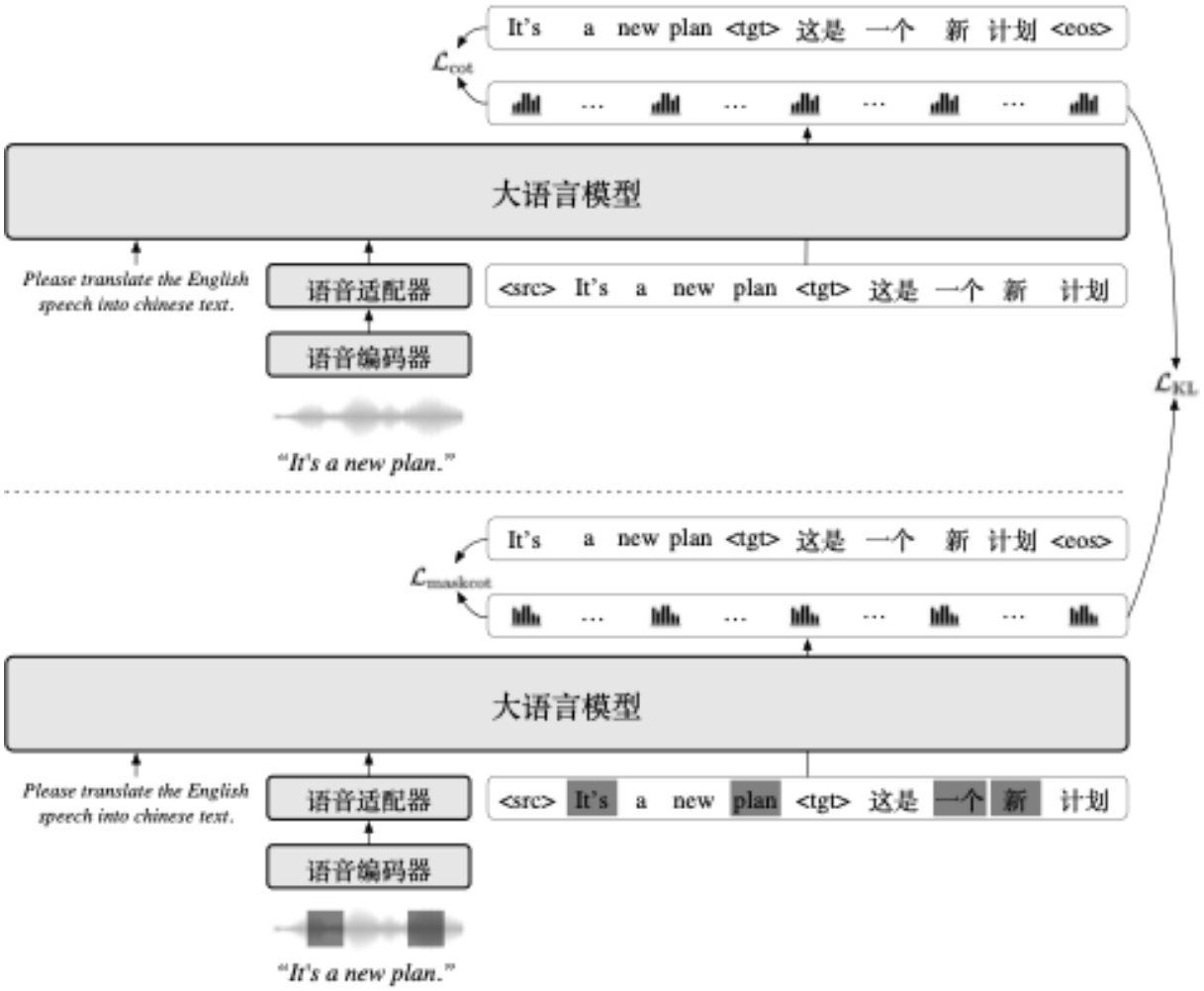
知识蒸馏（Knowledge Distillation）[45-47]：将文本翻译模型作为教师模型，利用其产生的译文或软标签来指导语音翻译学生模型的训练，从而进行有效的知识迁移；

数据增强（Data Augmentation）[48-51]：通过数据合成来扩充训练数据，例如，利用语音合成模型将文本翻译数据的源文本合成为语音，以此构造语音翻译的伪平行数据，从而扩充训练数据规模。

尽管上述方法在一定程度上缓解了数据问题，但语音与文本之间固有的“模态鸿沟”（Modality Gap）仍是充分利用机器翻译数据的主要障碍。为跨越这一鸿沟，研究者们探索了多种途径，例如将语音和文本特征投影到共享语义空间[17]，混合不同模态的特征以促使其学习相似的表示[24,52]，运用对比学习拉近句子级别的跨模态表征[53-54]，设计能够进行语音文本联合预训练的模型架构[41,55-58]，通过生成对抗训练来进行跨模态软对齐[59]等方法。此外，Deng 等人[60]提出引入语义解码器来生成与文本特征在分布和长度上更一致的语音特征。更进一步地，Zhu 等人[61]探索了视觉-音频-文本多模态联合预训练方法，通过统一的掩码预测任务学习跨模态的通用表征，以弥合不同模态间的隔阂。

随着以ChatGPT为代表的大语言模型的崛起，基于 LLM 的语音翻译展现出超越传统序列到序列模型的潜力。为充分发掘 LLM 在语音翻译中的能力，研究者们探索了多种技术方案。例如，Zhang 等人[62]利用前馈神经网络（Feed-Forward Network，简称 FFN）层连接预训练语音模型 Wav2vec 2.0 与 LLM，相比传统语音翻译模型达到了更好的性能。为解决语音与文本序列长度不匹配的问题，Wu 等人[27]采用了连接时序分类（Connectionist Temporal Classification，简称 CTC）对输入语音特征序列进行压缩，Wang 等人[28]通过随机丢弃大部分语音帧（如 75%）来实现长度压缩，Yu 等人[33]使用窗级别 Q-Former 结构进行语音特征的压缩与对齐。Huang等人[63]运用多任务学习策略，结合语音翻译、机器翻译、语音识别、标点恢复以及文本平滑等任务进行联合优化，以构建工业级性能的语音翻译系统。Chen 等人[64]探索了双 LoRA（Low-Rank Adaptation）优化、多语言增强训练等训练技术。Du 等人[34]和 Hu 等人[35]提出在语音翻译过程中融入思维链机制，通过依次生成源语言文本和目标语言文本来提升最终的翻译性能。Lam 等人[65]系统性地探究了特征拼接和跨注意力机制在语音大模型

图 2　模型结构与训练方法示意图



中的表现差异。Liu 等人[66]和 Zhang 等人[67]通过在大语言模型内部进行语音文本跨模态对齐来提高语音翻译的性能。

3 方法

3.1 任务定义

语音翻译是指将源语言的语音转换为目标语言的文本，其数据集通常采用三元组形式表示，记作$\mathcal{D} = (S, X, Y)$，其中$S = (s_1, \dots, s_{|S|})$表示源语言语音，$X = (x_1, \dots, x_{|X|})$表示对应的源语言文本，$Y = (y_1, \dots, y_{|Y|})$表示目标语言文本。

3.2 模型结构

本节将介绍本文所采用的模型结构。具体而言，整个模型主要由三部分组成：语音编码器（Speech Encoder）、语音适配器（Speech Adaptor）以及大语言模型，如图 2 所示。其中，语音编码器用于提取源语言语音的语义特征，语音适配器则负责对其进行长度压缩与特征映射，以转换为适配大语言模型输入的表示形式。最终，大语言模型根据转换后的表示生成对应的目标语言文本。

3.2.1 语音编码器

本文采用 Whisper-large-v3[31]的编码器作为语音编码器。Whisper 是一个在大规模多语言音频数据上训练得到的通用语音识别模型，其编码器具备强大的语义建模能力，能够有效捕捉语音中的语言信息与上下文特征。

Whisper 编码器整体结构由两部分组成：首先，源语言语音的对数梅尔频谱图（Log-Mel Spectrogram）经过两个一维卷积层进行初步特征提取与时间维度下采样；随后，得到的中间表示输入至若干 Transformer 编码器块中，进一步建模上下文表示。通过上述编码器的处理，输入语音最终被映射为一组高层次的语音特征表示，记为：

$$\boldsymbol{H} = (\boldsymbol{h}_1, \dots, \boldsymbol{h}_T),$$

其中，$\boldsymbol{h}_i \in \mathbb{R}^d$表示第$i$帧的语音特征向量，$T$为时间维度上的帧数，$d$为特征维度。

3.2.2 语音适配器

语音适配器的作用是将语音编码器输出的特

征$\boldsymbol{H}$转换为适配大语言模型输入的形式。具体而言，语音适配器包括两步处理：长度适配与特征适配。

在长度适配阶段，考虑到语音序列的时间维度远长于对应的文本序列，直接输入到大语言模型将带来较大的计算开销，且影响建模效果。例如，1秒钟的语音经过 Whisper 编码后产生 50 个特征帧，而对应的文本通常仅包含 3～5 个词。为此，本文采用下采样策略，将每k个连续的特征帧在特征维度上进行拼接，从而实现k倍的下采样。定义处理后的特征序列为$\boldsymbol{H}^{'}=\left[\boldsymbol{h}_1^{'},\ldots,\boldsymbol{h}_{\lfloor T/k\rfloor}^{'}\right]$，其中每个新的特征$\boldsymbol{h}_i^{'}$由$k$个连续帧拼接而成：

$$\boldsymbol{h}_i^{'}=\left[\boldsymbol{h}_{k\times(i-1)+1}\oplus\boldsymbol{h}_{k\times(i-1)+2}\oplus\cdots\oplus\boldsymbol{h}_{k\times i}\right].$$

在特征适配阶段，本文采用两层前馈神经网络将下采样后的特征表示映射到大语言模型的输入嵌入空间，具体计算如下：

$$\boldsymbol{S}=\mathrm{Linear}\left(\mathrm{ReLU}\left(\mathrm{Linear}(\boldsymbol{H}^{'})\right)\right),$$

其中，Linear表示线性变换层，ReLU表示线性整流激活函数，$\boldsymbol{S}$为语音适配器最终输出的特征序列，作为大语言模型的输入。

3.2.3 大语言模型

本文采用 Qwen2.5 系列的大语言模型[68]作为语音翻译的生成模块。Qwen2.5 是一种基于仅解码器架构的自回归语言模型，由堆叠的 Transformer 解码器块组成。其核心结构包括因果自注意力机制（Causal Self-Attention）、前馈神经网络、旋转位置编码（Rotary Positional Embedding，简称 RoPE）和残差连接（Residual Connection）等。与双向注意力不同，因果自注意力在每个位置上仅关注该位置及其之前的 token，从而符合自回归语言建模的假设。Qwen2.5 模型在大规模多语言文本数据上进行了预训练，具备强大的上下文建模能力和跨语言迁移能力，在纯文本机器翻译任务中表现优异。

为了将 Qwen2.5 应用于语音翻译任务，本文将提示词（Prompt）P的嵌入（Embedding）表示$\mathrm{Emb}(P)$与语音适配器输出的特征序列$\boldsymbol{S}$在时间维度上拼接作为输入：

$$\boldsymbol{Z}=[\mathrm{Emb}(P);\boldsymbol{S}],$$

其中，$\mathrm{Emb}(P)$表示通过嵌入层获得的提示词向量序列，$[\cdot;\cdot]$表示在时间维度上的拼接操作。对于语音翻译任务，提示词 $P=$ *"Please translate the {src_lang} speech into {tgt_lang} text."*。其中，*src_lang* 和 *tgt_lang* 分别代表源语言和目标语言。

基于上述输入，模型通过自回归方式逐 token 生成目标语言文本$Y=(y_1,y_2,\ldots,y_{|Y|})$，其训练目标为最小化负对数似然损失函数，形式如下：

$$\mathcal{L}_{\mathrm{st}}=-\sum_{i=1}^{|Y|}\log P\left(y_i\mid Y_{<i},\boldsymbol{Z}\right).$$

3.3 鲁棒思维链增强

为了进一步提升语音翻译的质量，本文在翻译过程中引入思维链（Chain-of-Thoughts）机制，引导模型在生成目标语言文本之前，先从源语音中生成对应的源语言文本。该机制通过显式建模中间转录结果，将转录与翻译过程解耦，使模型在生成译文时能够参考源语言文本，从而降低直接从语音生成目标语言文本的难度。与此同时，该方法仍保持端到端的训练流程，允许模型在翻译阶段同时利用原始语音信息与源语言文本。在该思维链建模策略下，本文对输入提示词P进行如下设置：$P=$ *"Please first transcribe the {src_lang} speech into text, and translate it into {tgt_lang}."*

在该设置下，模型的目标输出序列可表示为Y^{cot}，其形式为：

$$Y^{\mathrm{cot}}=\texttt{<src>}\,X\,\texttt{<tgt>}\,Y,$$

其中，<src>和<tgt>分别为源语言文本X与目标语言文本Y的起始标识符，模型采用自回归方式生成Y^{cot}，训练目标如下：

$$\mathcal{L}_{\mathrm{cot}}=-\sum_{i=1}^{|Y^{\mathrm{cot}}|}\log P\left(y_i^{\mathrm{cot}}\mid Y_{<i}^{\mathrm{cot}},\boldsymbol{Z}\right).$$

尽管思维链的建模方式通常能够显著提升翻译质量，但由于自回归生成的特性，模型在生成目标语言文本时可能过于依赖中间生成的源语言文本，从而忽略了原始语音信息。考虑到中间的源语言文本可能存在错误，这种建模方式易引发类似级联系统中的“错误传播”问题，即模型基于错误的源语言文本生成了不准确的目标语言翻译。为此，本文提出一种用于提升思维链鲁棒性的训练策略，旨在增强模型在面对存在错误的中间源语言文本时，仍能生成准确目标语言翻译的能力。

具体而言，在训练阶段，本文以一定的概率α对思维链部分的文本Y^{cot}进行随机掩码操作。该操作形式化表示如下：

$$\hat{Y}^{cot}=\mathcal{M}(Y^{\mathrm{cot}})=\left[\hat{y}_1^{cot},\ldots,\hat{y}_{|Y^{cot}|}^{cot}\right],$$

其中：

$$\hat{y}_i^{cot} = \begin{cases} 0, \text{ if } \ p < \alpha \\ y_i^{\text{cot}}, \text{ otherwise'} \end{cases}$$

p是从均匀分布$U(0,1)$中采样的随机变量。此外，为了增强模型对输入语音的鲁棒性，本文还在语音特征$\boldsymbol{S}$上也进行了随机掩码。具体地，对语音特征应用掩码操作得到：

$$\widehat{\boldsymbol{S}} = \mathcal{M}(\boldsymbol{S}),$$

模型的最终输入表示为：

$$\widehat{\boldsymbol{Z}} = [\text{Emb}(P); \widehat{\boldsymbol{S}}].$$

此时，模型需要基于掩码后的语音特征和思维链，按照思维链的方式依次生成源语言文本和目标语言文本，训练目标如下：

$$\mathcal{L}_{\text{maskcot}} = -\sum_{i=1}^{|Y^{\text{cot}}|} \log P\left(y_i^{\text{cot}} \mid \hat{Y}_{<i}^{cot}, \widehat{\boldsymbol{Z}}\right).$$

通过该训练策略，模型在面对不完整的思维链时仍需生成完整的目标译文，从而避免在训练过程中对源语言文本的过度依赖，增强其对源语言语音和文本的联合建模能力。因此，在推理阶段，即便思维链中存在错误，模型仍将具备一定的鲁棒性。

由于模型在解码阶段生成的思维链不再进行随机掩码，若仅优化$\mathcal{L}_{\text{maskcot}}$，将导致训练与推理阶段存在分布不一致的问题。为缓解该问题，本文在训练阶段联合优化$\mathcal{L}_{\text{maskcot}}$与$\mathcal{L}_{\text{cot}}$，即模型需在完整与不完整思维链条件下均能生成正确的译文。同时，为防止模型在优化$\mathcal{L}_{\text{cot}}$时过度依赖完整思维链，本文引入 KL 散度（Kullback-Leibler Divergence）进行约束，使两种条件下的预测分布保持一致性，其优化目标为：

$$\mathcal{L}_{\text{KL}} = -\sum_{i=1}^{|Y^{\text{cot}}|} \text{KL}\left(P(y_i^{\text{cot}} \mid Y_{<i}^{\text{cot}}, \boldsymbol{Z}) || P\left(y_i^{\text{cot}} \mid \widehat{Y_{<i}^{\text{cot}}}, \widehat{\boldsymbol{Z}}\right)\right).$$

最终，模型的优化目标如下：

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{cot}} + \mathcal{L}_{\text{maskcot}} + \mathcal{L}_{\text{KL}}.$$

上述训练方法如图 2 所示。

3.4 模型训练

本文的模型训练采用单阶段策略。在训练过程中，语音编码器参数保持冻结，以保留 Whisper 模型的预训练表示能力，避免微调过程中其泛化能力的丢失。语音适配器的全部参数参与训练。

对于大语言模型部分，为提高模型训练效率，本文在模型的 FFN 模块中引入 LoRA（Low-Rank Adaptation）进行高效模型微调。具体而言，LoRA 应用于 FFN 的 gate_proj、up_proj 与 down_proj 三个线性层中，训练过程中仅优化新增的低秩参数，保持大语言模型的主体参数冻结，从而显著降低训练开销。

表 1 CoVoST 2 六个语向的数据统计（单位：小时）

语向（X→En）	训练集	开发集	测试集
法语（Fr）	180	22	23
德语（De）	119	21	22
西班牙语（Es）	97	22	23
意大利语（It）	28	14	15
中文（Zh）	10	8	8
日语（Ja）	1	1	1

4 实验

4.1 数据集

本文采用 CoVoST 2 数据集[69]进行实验。CoVoST 2 是一个大规模多语言语音翻译语料库，涵盖了 21 种语言翻译至英语以及英语翻译至 15 种语言的数据。遵循 Chen 等人[64]的实验设置，本文选取了从法语（Fr）、德语（De）、西班牙语（Es）、意大利语（It）、中文（Zh）及日语（Ja）翻译至英语的六个语向。表 1 展示了这六个语向在训练集、开发集和测试集上的数据统计。在训练阶段，本文将这六个语向的数据合并进行统一的多语言模型训练。最终，模型性能在上述六个语向各自的测试集上进行评估。在所有实验中，源语言语音均从 48kHz 下采样至 16kHz 作为模型输入。

此外，本文还在常用的语音翻译基准数据集 MuST-C[70]的英语到德语（En→De）方向上进行了补充实验。该实验的模型在 MuST-C En→De 的训练集上进行训练，其包含约 408 小时语音翻译数据，并在 tst-COMMON 测试集上进行评估。

4.2 模型配置

本文采用 Whisper-large-v3 的编码器作为语音编码器，语音适配器首先进行$k = 5$倍下采样，然后

表 2 不同模型在 CoVoST 2 六个翻译方向（X→En）上的 BLEU 分数

模型	Fr→En	De→En	Es→En	Zh→En	Ja→En	It→En	平均值
SpeechLLaMA	25.20	27.10	27.90	12.30	19.90	25.90	23.05
Whisper-large-v2	36.40	36.30	40.10	18.00	26.10	30.90	31.30
Whisper-large-v3	35.53	34.18	39.26	13.15	23.04	35.94	30.18
SeamlessM4T-large-v2	42.10	39.90	42.90	22.20	23.80	40.00	35.15
Qwen2-Audio	38.50	35.20	40.00	24.40	N/A	36.30	N/A
LLaST-8B	44.10	40.80	45.30	23.30	24.40	42.10	36.67
Ours-ST-7B	41.99	39.65	44.07	21.99	23.52	41.23	35.41
Ours-CoT-7B	42.86	40.33	44.43	27.36	26.94	41.70	37.27
Ours-RobustCoT-7B	**44.21**	**41.48**	**45.46**	**27.81**	**27.11**	**43.08**	**38.19**

经过中间维度为 2048 的 2 层 FFN 进行特征变换。对于大语言模型部分，本文采用 Qwen2.5 系列模型作为大语言模型，包含 Qwen2.5-0.5B/1.5B/3B/7B-Instruct 四个不同尺寸的模型。

4.3 模型训练

在训练阶段，语音编码器的参数均予以冻结，而语音适配器中 FFN 的参数则进行训练。对于大语言模型，本文为所有 FFN 模块中的三个线性层添加 LoRA 参数，其中 LoRA 的秩（rank）设为 512，学习率缩放因子设为 1024，dropout 率设为 0.05。优化器选用 AdamW[71]，其中$\beta_1 = 0.9, \beta_2 = 0.999$。训练的批处理大小（batch size）为 32，模型在 CoVoST 2 六个语向合并的训练数据上训练 1 个轮次（epoch）。学习率在前 3%的训练步数内从 0 线性预热（warmup）至10^{-4}，随后采用余弦退火（cosine annealing）策略进行衰减。对于训练中引入的随机掩码操作，掩码概率设为$\alpha = 0.2$，KL 散度的权重系数设为$\lambda = 1.0$。所有模型的训练均在配备 4 块 NVIDIA H800 GPU 的服务器上完成。

4.4 模型评估

在评估阶段，模型采用束搜索（beam search）进行解码，束宽（beam size）设置为 5。翻译性能的主要评估指标为 BLEU（Bilingual Evaluation Understudy）[72]。该指标通过比较机器翻译输出与人工参考翻译在 n-gram 层面的重叠度，并结合对输出长度的惩罚机制，来评价翻译质量。具体的 BLEU 分数由 SacreBLEU 工具包[73]计算得出。

4.5 基线系统

为全面评估本文所提出模型的性能，本文选取了语音翻译领域多个具有代表性的模型作为基线系统。这些模型涵盖了不同的架构和技术方案，其简要介绍如下：

表 3 MuST-C 英语→德语方向的 BLEU 分数

模型	BLEU
SeamlessM4T-large-v2	25.35
SD-ST	27.20
SD-ST-Large	27.90
Qwen2-Audio	30.23
Ours-ST-7B	30.59
Ours-CoT-7B	31.68
Ours-RobustCoT-7B	**32.45**

SpeechLLaMA[27]：SpeechLLaMA 首先采用一个基于 CTC 的压缩模块对声学特征进行初步处理，随后通过音频编码器提取深层语音表征。语音表征与文本提示的嵌入向量拼接后输入大语言模型，从而生成目标语言文本。

Whisper[31]：Whisper 系列模型是由 OpenAI 开发的端到端语音识别与翻译模型。该系列模型在包含约 68 万小时的多语言、多任务语音数据集上进行训练，取得了优异的语音识别与翻译性能。本文选取了该系列中的 Whisper-large-v2 和 Whisper-large-v3 模型作为基线系统。

Qwen2-Audio[74]：Qwen2-Audio 是由阿里巴巴通义千问团队推出的一个通用音频理解大模型。该模型基于 Qwen-7B 语言模型进行构建，能够接受各种音频信号输入，并根据指令执行音频分析或直接响应文本。

SeamlessM4T[75]：SeamlessM4T 是由 Meta AI 推出的一个大规模多语言多模态的“一体化”翻译模型。该模型能够利用单一模型完成包括语音到语

表 4 不同参数规模模型在 CoVoST 2 六个翻译方向（X→En）上的 BLEU 分数

模型	Fr→En	De→En	Es→En	Zh→En	Ja→En	It→En	平均值
Ours-ST-0.5B	36.51	32.23	38.52	14.24	**19.39**	35.20	29.35
Ours-CoT-0.5B	38.51	33.62	40.18	18.61	16.85	36.11	30.65
Ours-RobustCoT-0.5B	**39.17**	**34.84**	**40.39**	**18.62**	18.96	**36.90**	**31.48**
Ours-ST-1.5B	39.42	35.47	41.54	16.96	20.44	38.17	32.00
Ours-CoT-1.5B	40.94	37.05	42.07	**23.34**	21.05	39.20	33.94
Ours-RobustCoT-1.5B	**41.95**	**38.33**	**43.18**	22.59	**22.74**	**40.21**	**34.83**
Ours-ST-3B	40.61	37.96	43.19	19.96	20.73	39.83	33.71
Ours-CoT-3B	41.83	38.95	43.62	**25.17**	22.69	41.12	35.56
Ours-RobustCoT-3B	**43.22**	**40.16**	**44.46**	**25.17**	**22.95**	**41.84**	**36.30**
Ours-ST-7B	41.99	39.65	44.07	21.99	23.52	41.23	35.41
Ours-CoT-7B	42.86	40.33	44.43	27.36	26.94	41.70	37.27
Ours-RobustCoT-7B	**44.21**	**41.48**	**45.46**	**27.81**	**27.11**	**43.08**	**38.19**

音、语音到文本、文本到语音以及文本到文本在内的多种翻译与转录任务，并支持近百种语言的输入与输出。本文选取其公开的 SeamlessM4T-large-v2 模型作为基线系统。

SD-ST[60]：SD-ST 是一种基于语义解码器的语音翻译模型，该模型通过生成富含语义信息的语音特征来改善端到端的语音到文本翻译。本文选取 SD-ST 和 SD-ST-Large 模型作为基线系统。

LLaST[64]：LLaST 是一个结合语音编码器与大语言模型的语音翻译框架，针对语音翻译场景提出了 ASR 增强训练、多语言数据增强以及双 LoRA 优化策略。通过这些改进，LLaST 在 CoVoST 2 等基准测试上展现了优越的性能，本文选取 LLaST-8B 模型作为基线系统。

除了上述已有的基线系统外，为了验证本文所提出方法的有效性并进行充分的比较与分析，本文构建并评估了以下三个不同版本的模型：

Ours-ST：这是本文所构建系统的基础版本。该模型直接利用语音翻译任务进行端到端训练，旨在建立一个初步的性能基准。

Ours-CoT：在 Ours-ST 模型的基础上，本文引入了常规的思维链增强训练方法。通过引导模型逐步生成源语言文本和目标语言文本，旨在提高模型的翻译准确性。

Ours-RobustCoT：为进一步提升模型应对思维链中潜在错误的能力，在 Ours-CoT 的基础上，本文融入了鲁棒思维链增强训练策略（如 3.3 节所述）。该策略通过训练阶段添加随机掩码以及对模型预测分布添加正则约束，旨在缓解模型过度依赖思维链导致的错误传播问题，从而获得更可靠的翻译结果。该模型代表了本文最终提出的优化方案。

对于 Ours-ST、Ours-CoT 及 Ours-RobustCoT 模型，本文选用 Qwen2.5 系列不同参数规模的模型作为基础分别构建了相应模型。在主实验中，本文重点采用 7B 参数规模的版本（相应命名为 Ours-ST-7B、Ours-CoT-7B 与 Ours-RobustCoT-7B）与现有基线系统进行性能对比。此外，在后续的分析实验中，还将对这些模型在不同参数规模下的性能表现进行比较研究。

5 结果与分析

5.1 实验结果

本文在 CoVoST 2 六个翻译方向（X→En）上进行了实验，旨在将本文提出的模型与多个主流基线系统进行语音翻译性能的综合比较。实验结果如表 2 所示，本文提出的 Ours-RobustCoT 模型在全部六个语向上均取得了最佳性能，其平均 BLEU 分数为 38.19，相较于此前性能最优的 LLaST-8B 模型提升了 1.52 BLEU 分，初步证实了本文方法的整体有效性。

具体分析本文提出的模型系列：首先，基线模型 Ours-ST-7B 的性能所展现的性能已能与业界主流的 SeamlessM4T-large-v2 模型相媲美。值得注意的是，Ours-ST-7B 仅使用了数百小时的训练数据便达到了此效果，这体现了基于大语言模型构建语音

表 5 不同训练目标组合下 Ours-RobustCoT-7B 模型的 BLEU 分数

$\mathcal{L}_{\text{cot}}$	$\mathcal{L}_{\text{maskcot}}$	$\mathcal{L}_{\text{KL}}$	Fr→En	De→En	Es→En	Zh→En	Ja→En	It→En	平均值
✓	✓	✓	**44.21**	**41.48**	**45.46**	**27.81**	**27.11**	**43.08**	**38.19**
✓	×	✓	43.06	40.52	44.36	25.28	25.51	41.86	36.77
✓	✓	×	42.53	40.16	44.49	27.62	26.48	41.29	37.10
✓	×	×	42.86	40.33	44.43	27.36	26.94	41.70	37.27

表 6 不同随机掩码位置下 Ours-RobustCoT-7B 模型的 BLEU 分数

语音	思维链	Fr→En	De→En	Es→En	Zh→En	Ja→En	It→En	平均值
✓	✓	**44.21**	41.48	**45.46**	27.81	**27.11**	**43.08**	**38.19**
×	✓	44.07	**41.54**	**45.46**	**28.35**	25.71	42.99	38.02
✓	×	43.18	40.74	44.82	27.91	26.52	42.19	37.56
Ours-CoT-7B		42.86	40.33	44.43	27.36	26.94	41.70	37.27

翻译系统的潜力与优势。其次，在 Ours-ST-7B 的基础上引入常规思维链训练方法后，Ours-CoT-7B 的平均 BLEU 分数提升了 1.86 分，证明思维链不仅显著增强了整体语音翻译性能，其效果在训练数据相对匮乏的翻译方向（例如中文→英文、日语→英文）上表现得尤为突出。更进一步，采用本文提出的鲁棒思维链训练方法，Ours-RobustCoT-7B 模型在 Ours-CoT-7B 的基础上进一步实现了 0.92 BLEU 分的提升，相较于初始基线模型 Ours-ST-7B，累计提升高达 2.78 BLEU 分。这些结果充分展现了本文所提出的鲁棒思维链方法在提升语音翻译质量方面的优越性。

为进一步验证模型的有效性，本文在 MuST-C 英语→德语数据集上进行了对比实验。如表 3 所示，本文提出的基线模型 Ours-ST-7B 取得了 30.59 的 BLEU 分数，优于 SeamlessM4T-large-v2、SD-ST、Qwen2-Audio 等多个基线模型。在引入思维链后，Ours-CoT-7B 模型性能提升至 31.68。进一步引入本文提出的鲁棒思维链训练策略后，Ours-RobustCoT-7B 模型进一步取得了 0.77 BLEU 分数的提升，进一步验证了本文所提出的鲁棒思维链方法的有效性。

5.2 不同参数规模模型的结果

为进一步探究模型参数规模对本文所提方法性能的影响，本文在 CoVoST 2 数据集的六个 X→En 翻译方向上，对不同参数规模（0.5B、1.5B、3B 及 7B）的 Ours-ST、Ours-CoT 以及 Ours-RobustCoT 模型进行了一系列对比实验。详细实验结果如表 4 所示。

整体而言，实验结果清晰地表明，对于本文提出的三种模型变体（Ours-ST、Ours-CoT、Ours-RobustCoT），其平均 BLEU 分数均随着模型参数规模的增大而稳步提升。以本文最终提出的 Ours-RobustCoT 模型为例，其平均 BLEU 分数从 0.5B 参数规模的 31.48 分逐步增长至 7B 参数规模的 38.19 分，说明模型的参数规模对最终的翻译质量起到决定性的作用。此外，在各个参数规模下，引入常规思维链训练的 Ours-CoT 模型均一致地优于对应的 Ours-ST 基线模型。在 0.5B、1.5B、3B 和 7B 规模下，分别带来了 1.30、1.94、1.85 和 1.86 BLEU 分的提升，对于中文→英文和日语→英文方向提升尤其明显，这证明了 CoT 策略在不同参数规模下均具有普适的提升效果。更进一步，本文提出的鲁棒思维链方法，在所有参数规模中，相比常规 CoT 均带来了进一步的提升。具体而言，Ours-RobustCoT 相较于 Ours-CoT 模型，在 0.5B、1.5B、3B 和 7B 规模下分别带来了约 0.83、0.89、0.74 和 0.92 BLEU 分的额外平均性能增益。这些结果充分表明，本文提出的鲁棒思维链方法不仅在较大模型（如 7B）上表现卓越，其优势同样能够推广并体现在参数规模较小的模型上，具有良好的一致性和扩展性。同时，实验结果也印证了增加模型参数规模是提升基于大语言模型的语音翻译系统性能的最有效途径之一。

5.3 不同训练目标组合下的结果

本文的训练目标由三个核心组成部分构成：基于完整思维链翻译的损失$\mathcal{L}_{\text{cot}}$、基于带掩码（不完整）思维链翻译的损失$\mathcal{L}_{\text{maskcot}}$，以及一个 KL 散度约束项$\mathcal{L}_{\text{KL}}$。为深入探究各训练目标及其不同组合对最终翻译性能的具体贡献，本文进行了一系列消

表 7 不同随机掩码概率α下 Ours-RobustCoT-7B 模型的 BLEU 分数

模型	Fr→En	De→En	Es→En	Zh→En	Ja→En	It→En	平均值
Ours-CoT-7B	42.86	40.33	44.43	27.36	26.94	41.70	37.27
Ours-RobustCoT-7B ($\alpha = 0.1$)	44.01	41.43	45.44	**28.00**	26.44	42.87	38.03
Ours-RobustCoT-7B ($\alpha = 0.2$)	**44.21**	41.48	**45.46**	27.81	**27.11**	43.08	**38.19**
Ours-RobustCoT-7B ($\alpha = 0.3$)	44.03	**41.66**	45.36	26.65	25.87	**43.10**	37.78
Ours-RobustCoT-7B ($\alpha = 0.4$)	42.97	40.69	44.18	24.71	26.44	41.68	36.78
Ours-RobustCoT-7B ($\alpha = 0.5$)	43.25	40.84	44.58	21.93	24.73	42.09	36.24

表 8 思维链中的语音识别错误率（%）

模型	Fr	De	Es	Zh	Ja	It	平均值
Ours-CoT-7B	9.30	6.48	4.80	12.34	18.88	6.18	9.66
Ours-RobustCoT-7B	**8.78**	**6.00**	**4.58**	**11.78**	**17.99**	**5.81**	**9.16**

融实验，结果详见表 5。

实验观察表明，以仅包含$\mathcal{L}_{\text{cot}}$的常规思维链训练为参照（平均 BLEU 分数为 37.27）：

1. 当在$\mathcal{L}_{\text{cot}}$基础上额外引入$\mathcal{L}_{\text{maskcot}}$（即$\mathcal{L}_{\text{cot}} + \mathcal{L}_{\text{maskcot}}$）时，模型平均 BLEU 分数出现了微弱下降，从 37.27 降至 37.10。本文推测，这可能是因为模型虽然学习了依据不完整思维链进行翻译的能力，但在面对完整思维链输入时，仍可能表现出一定程度的过度依赖，导致整体性能未获显著提升甚至略有回落。

2. 若仅在$\mathcal{L}_{\text{cot}}$基础上添加$\mathcal{L}_{\text{KL}}$损失项（即$\mathcal{L}_{\text{cot}} + \mathcal{L}_{\text{KL}}$），模型性能同样呈现下降趋势，平均 BLEU 分数从 37.27 降至 36.77。此现象的原因可能在于，此时模型未能对不完整思维链场景下的翻译提供有效监督，因此$\mathcal{L}_{\text{KL}}$对完整思维链预测分布施加的约束反而可能引致模型学习方向的潜在偏差。

3. 然而，当同时引入$\mathcal{L}_{\text{maskcot}}$和$\mathcal{L}_{\text{KL}}$的完整训练目标时，模型性能获得了显著提升，平均 BLEU 分数从 37.27 提升至 38.19。此时，模型得以同时学习在完整及不完整思维链条件下进行翻译的能力，并且$\mathcal{L}_{\text{KL}}$有效约束了两种情况下的预测分布，减轻了模型对思维链的过度依赖，从而带来了最终性能的明显进步。

综上所述，尽管单独引入$\mathcal{L}_{\text{maskcot}}$或$\mathcal{L}_{\text{KL}}$可能无法直接提升甚至会小幅影响基于$\mathcal{L}_{\text{cot}}$的基线性能，但三者的结合能够有效提高翻译性能，证明了本文所设计的训练方法的有效性。

5.4 不同随机掩码位置下的结果

本文进一步探究了在不同输入位置施加随机掩码对模型性能的影响。在本文提出的 Ours-RobustCoT 模型中，采用了在语音输入和思维链输出两部分均施加随机掩码的策略，具体而言，是以$\alpha = 0.2$的概率随机将部分 token 替换为零值。为分别评估在语音输入和思维链输出上施加随机掩码的独立及联合作用，本文进行了一系列消融实验，其结果详见表 6。

实验结果显示，以不包含掩码的 Ours-CoT-7B 模型作为基准（平均 BLEU 分数为 37.27）：

1. 当仅对语音输入施加掩码时，模型性能相较于 Ours-CoT-7B 实现了平均 0.29 BLEU 的微小提升。此增益可主要归因于模型对输入语音扰动的鲁棒性得到了一定增强。

2. 相比之下，若仅对思维链输出施加掩码，模型的平均 BLEU 分数较 Ours-CoT-7B 基线则有平均 0.75 BLEU 的显著提升。这表明在思维链上进行掩码操作，能够有效缓解模型对所生成思维链的过度依赖问题。

3. 最终，同时在语音输入与思维链输出两处施加掩码时，模型性能获得了进一步的提升，达到了 38.19 的平均 BLEU 分数，相较于 Ours-CoT-7B 提升了 0.92 BLEU 分。

上述实验结果表明，尽管主要的性能提升源自于对思维链部分施加的随机掩码，但对语音输入进行掩码同样能带来额外的性能增益。因此，二者结合的掩码方案对于模型实现最优性能是必要的。

5.5 不同随机掩码概率α下的结果

为进一步探究随机掩码策略中掩码概率α对模型性能的具体影响，本文对 Ours-RobustCoT-7B 模型在不同α取值下的表现进行了细致评估。表 7 结果清晰地揭示了模型翻译性能随α值变化的趋势。

表 9 级联系统在 CoVoST 2 六个翻译方向（X→En）上的 BLEU 分数

模型	Fr→En	De→En	Es→En	Zh→En	Ja→En	It→En	平均值
Whisper-large-v3 + Qwen2.5-7B-Instruct	36.96	36.47	39.95	19.10	25.63	36.47	32.43
Whisper-large-v3 + Qwen2.5-7B-Finetune	38.94	39.24	42.92	24.09	**27.12**	39.29	35.27
Ours-ST-7B	41.99	39.65	44.07	21.99	23.52	41.23	35.41
Ours-CoT-7B	42.86	40.33	44.43	27.36	26.94	41.70	37.27
Ours-RobustCoT-7B	**44.21**	**41.48**	**45.46**	**27.81**	27.11	**43.08**	**38.19**

具体而言，随着α值从 0 开始增加，模型的平均 BLEU 分数起初呈现稳步提升的态势，并在$\alpha = 0.2$时达到了 38.19 BLEU 的峰值。此后，若α值继续增大，模型性能则表现出逐渐下降的趋势。这一“先升后降”的性能曲线表明，适当的随机掩码能够有效提升模型的翻译性能，增强模型的鲁棒性。然而，过度的掩码（即α值过高）则会引入过强的噪声，使得训练任务过于困难，反而导致性能显著下降。因此，选择合适的掩码概率对于最终的翻译性能至关重要。综合实验结果，本文最终选择$\alpha = 0.2$作为 Ours-RobustCoT 模型中随机掩码的概率。

5.6 思维链中的语音识别错误率

由于思维链的中间步骤包含了对源语言的语音转写，其准确性是影响后续翻译质量的关键因素之一。因此，本文对思维链中间过程生成的源语言转写文本的错误率进行了评估。具体而言，本文针对法语（Fr）、德语（De）、西班牙语（Es）及意大利语（It）的评测采用了词错误率（Word Error Rate，简称 WER），而针对中文（Zh）和日语（Ja）则采用更适合其语言特性的字符错误率（Character Error Rate，简称 CER）作为评价指标。

如表 8 所示，本文提出的鲁棒思维链模型 Ours-RobustCoT-7B 相较于常规的思维链模型 Ours-CoT-7B，在所有六个语言上的语音识别错误率均有所降低。总体来看，其平均错误率由 9.66% 下降至 9.16%，获得了 0.5 个百分点的改善。上述结果证明，本文所提出的鲁棒思维链方法不仅能提升最终的翻译质量，也能有效增强对源语言的语音识别准确率，从而为翻译过程提供更加可靠的中间步骤。

5.7 对语音识别错误的鲁棒性分析

为进一步量化验证本文所提出的鲁棒思维链方法在面对不完美中间转录时的有效性，本文对模型在不同语音识别错误率下的翻译性能进行了分组分析。本文将测试集根据 Ours-CoT-7B 模型生成的思维链中间转录文本的错误率（WER 或 CER）进行划分，并对比了 Ours-CoT-7B 与 Ours-RobustCoT-7B 模型在这些不同错误率区间的 BLEU 分数，结果如图 3 所示。如图所示，本文所提出的 Ours-RobustCoT-7B 模型在几乎所有翻译方向和错误率区间上，性能均优于或持平于常规的 Ours-CoT-7B 模型。尤为关键的是，随着中间转录文本的错误率上升，本文方法的性能优势愈发显著，表明思维链中的转录错误越多，本文所提方法的改进效果越明显。以法语到英语（Fr→En）的翻译任务为例，当语音转录质量较高时（WER 0-20%），本文方法带来的 BLEU 值提升仅为 1.04 分；然而，当转录错误非常严重时（WER 80-100%），性能优势则达到了 5.76 BLEU 分，其他翻译方向上也展现出类似的趋势。这一现象清晰地表明，当中间转录出现较多错误时，常规思维链模型因过度依赖错误的文本信息而性能急剧下降，而本文提出的方法能够更有效地结合原始语音信号来纠正或忽略思维链中的错误，显著增强了模型在面对不完美转录时的鲁棒性。与此同时，从模型的绝对性能来看，尽管 Ours-RobustCoT-7B 表现更优，其翻译质量仍随着转录错误率的上升而下降。鉴于本文所采用的语音编码器 Whisper 本身已具备较强的鲁棒性，本文分析认为，极高的转录错误率可能是由于原始语音信号本身可辨识度极低（如高环境噪音、发音含糊不清等），而非简单的模型识别能力局限。在此条件下，模型即便具备一定的鲁棒性，也难以从根本上解决由源端信息失真所带来的翻译挑战。

5.8 与级联系统的结果对比

为明确对错误源文本的纠错能力是源于大语言模型固有的性能，还是本文提出的鲁棒思维链训练策略，本文进一步引入了两个级联系统进行对比。第一个系统名为 Whisper-large-v3+Qwen2.5-7B-Instruct，其首先使用 Whisper-large-v3 进行语音识别，而后通过提示的方式使用 Qwen2.5-7B-Instruct 模型进行翻译。第二个系统名为 Whisper-large-v3+

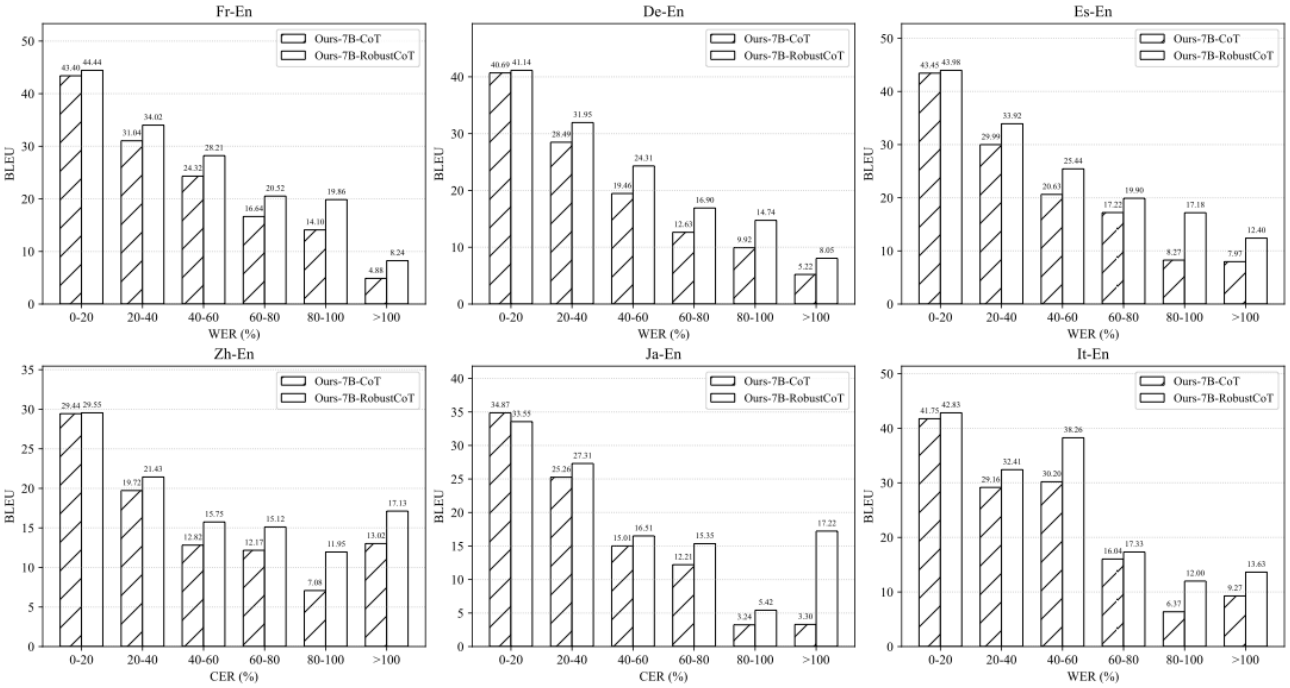


图 3　不同语音识别错误率区间的 BLEU 分数对比

表 10　翻译案例

示例编号	源语音对应文本	参考译文	Ours-ST-7B	Ours-CoT-7B	Ours-RobustCoT-7B
common_voice_zh-CN_18986599	藏书八万卷。	There are 80000 volumes of books.	Wrote over 80,000 words.	<src>**长书**八万卷。<tgt> **Long book** of 80,000 volumes.	<src>**长书**八万卷。<tgt> 80,000 volumes of books.
common_voice_zh-CN_18980424	曾祖父达文。	Grandfather Da Wen.	Grandma Da Wen.	<src>曾祖父**答文**。<tgt> His great-grandfather **answered**.	<src>曾祖父**答文**。<tgt> Great-Grandfather Da Wen.

Qwen2.5-7B-Finetune，其语音识别模块与前者相同，但翻译模块基于 Qwen2.5-7B-Instruct 在 CoVoST 2 的文本翻译数据上进行了微调，以对齐领域知识。如表 9 所示，实验结果表明，Whisper-large-v3+Qwen2.5-7B-Instruct 系统的性能显著低于本文的 Ours-ST-7B 基线模型。在对翻译模型进行同领域微调后，Whisper-large-v3+Qwen2.5-7B-Finetune 系统的性能有所提升，其平均 BLEU 分数与 Ours-ST-7B 模型基本持平，但与 Ours-RobustCoT-7B 模型相比仍存在显著差距。这一系列对比说明，大语言模型固有的纠错能力不足以完全弥补级联架构中错误传播带来的负面影响，其性能上限仅能与基础的端到端语音翻译模型相当。相比之下，思维链的引入能够有效提升端到端模型的性能，而本文提出的鲁棒思维链训练策略，通过在训练中强制模型结合原始语音信号进行判断，能进一步降低错误源文本的干扰，从而带来额外的性能增益。

5.9　案例分析

为从定性角度检验不同模型处理语音转录错误的能力，本节选取了两个典型案例进行分析，如表 10 所示。在基于常规思维链的模型 Ours-CoT-7B 中，当源语音的“藏”被错误转录为近音词“长”，或人名“达文”被转录为同音词“答文”时，出现了错误传播现象。其最终翻译（如“Long book...”和“...answered.”）直接反映了这些转录错误，并与原始语义不符。

相比之下，基于鲁棒思维链的模型 Ours-RobustCoT 在面对相同的转录错误时，表现出不同的行为。尽管其中间步骤生成了同样不准确的文本（“长书八万卷”和“曾祖父答文”），但其最终输出的翻译（“80,000 volumes of books.”和“Great-Grandfather Da Wen.”）与原始语音的语义保持了一致，并未受到中间文本错误的影响。

以上定性分析结果表明，本文提出的鲁棒性训

表 11　CoVoST 2 数据集 21 个翻译方向上的 BLEU 分数

模型	Ar→En	Ca→En	Cy→En	De→En	Es→En	Et→En	Fa→En

Ours-ST-7B-ML	39.54	35.23	12.43	38.06	43.54	16.05	14.95
Ours-CoT-7B-ML	44.77	36.32	**22.97**	40.55	44.99	19.18	17.96
Ours-RobustCoT-7B-ML	**46.00**	**38.67**	21.96	**41.58**	**45.76**	**19.65**	**21.33**
	Fr→En	Id→En	It→En	Ja→En	Lv→En	Mn→En	Nl→En
Ours-ST-7B-ML	41.06	50.66	40.06	22.08	18.11	0.53	39.68
Ours-CoT-7B-ML	43.09	56.35	42.10	**26.57**	25.16	2.82	44.38
Ours-RobustCoT-7B-ML	**44.16**	**56.88**	**43.41**	26.41	**25.81**	**2.95**	**44.97**
	Pt→En	Ru→En	Sl→En	Sv→En	Ta→En	Tr→En	Zh→En
Ours-ST-7B-ML	52.96	50.25	26.53	43.78	2.80	30.99	20.44
Ours-CoT-7B-ML	55.54	52.53	**34.93**	48.03	**3.81**	33.00	27.42
Ours-RobustCoT-7B-ML	**56.22**	**53.12**	34.61	**49.36**	3.76	**35.88**	**28.29**

表 12 不同模型在 Europarl-ST v1.1 四个翻译方向（X→En）上的 BLEU 分数

模型	Fr→En	De→En	Es→En	It→En	平均值
Ours-ST-7B	31.66	25.41	30.91	28.18	29.04
Ours-CoT-7B	32.41	26.13	31.13	28.55	29.56
Ours-RobustCoT-7B	**33.18**	**26.88**	**31.51**	**29.32**	**30.22**

练方法降低了模型对中间思维链文本的依赖。该机制促使模型在生成最终翻译时，结合利用原始语音信号与中间生成的文本，这提升了模型在面对不完美转录文本时，生成符合原始语义的翻译的能力，增强翻译结果的准确性与鲁棒性。

5.10 更多语向上的实验结果

为了进一步验证本文所提出方法在更多语言方向上的通用性，本文随后在 CoVoST 2 数据集所涵盖的全部 21 个“多对一”（X→En）语言方向上进行了更为全面的实验。这些语言方向具体包括：阿拉伯语 (Ar)、加泰罗尼亚语 (Ca)、威尔士语 (Cy)、德语 (De)、西班牙语 (Es)、爱沙尼亚语 (Et)、波斯语 (Fa)、法语 (Fr)、印度尼西亚语 (Id)、意大利语 (It)、日语 (Ja)、拉脱维亚语 (Lv)、蒙古语 (Mn)、荷兰语 (Nl)、葡萄牙语 (Pt)、俄语 (Ru)、斯洛文尼亚语 (Sl)、瑞典语 (Sv)、泰米尔语 (Ta)、土耳其语 (Tr) 以及中文 (Zh) 到英语的翻译。值得注意的是，CoVoST 2 数据集的语言方向覆盖了悬殊的数据资源规模，训练数据量从数百小时的高资源语种（如德语、法语）到仅约一小时的极低资源语种（如威尔士语、日语）不等，因此能够检验所提方法在不同资源约束下的性能和通用性。

在实验设置上，本文结合上述 21 个语向的全部训练数据，训练统一的多语言语音翻译模型。本文将语音翻译基准模型、基于常规思维链的模型和基于鲁棒思维链的模型分别记作 Ours-ST-7B-ML、Ours-CoT-7B-ML 和 Ours-RobustCoT-7B-ML，实验结果如表 11 所示。可以清晰地观察到，在所有翻译方向上，引入思维链的 Ours-CoT-7B-ML 模型相比传统的基准模型 Ours-ST-7B-ML 均取得了显著的性能提升。更进一步地，本文提出的鲁棒思维链方法 (Ours-RobustCoT-7B-ML) 在此基础上，于大多数语言方向上都获得了稳定的性能增益，展现了其优越性。尤其值得注意的是，在某些翻译方向上，该方法的提升十分显著。例如，在加泰罗尼亚语到英语 (Ca-En)、波斯语到英语 (Fa-En) 和土耳其语到英语 (Tr-En) 的任务中，本文所提出的鲁棒思维链模型相较于常规思维链模型，BLEU 值分别提升了 2.35、3.37 和 2.88。对于所有 21 个语向，Ours-CoT-7B-ML 相较于基准模型平均提升了 3.94 BLEU，而 Ours-RobustCoT-7B-ML 则在此基础上获得了额外的 0.88 BLEU 平均增益，系统性地验证了所提方法的总体有效性。上述实验结果证明本文所提出的方法在不同语言以及不同数据资源条件下均具备通用性，能够有效提升语音翻译鲁棒性。

5.11 泛化能力分析

为评估模型在训练领域之外的泛化能力，本文在域外数据集 Europarl-ST v1.1[76]上进行了补充测试：首先，模型在 CoVoST 2 数据集的六个翻译方向上完成训练；随后，选取与 Europarl-ST v1.1 测试集相重合的四个翻译方向进行测试。Europarl-ST v1.1 源自欧洲议会的正式会议记录，其主题领域与语言风格均与训练数据 CoVoST 2 存在显著差异，因此可有效检验模型的泛化能力。

如表 12 所示，在 Europarl-ST v1.1 测试集上，本文提出的三种模型保持了与 CoVoST 2 测试集上一致的相对性能排序。基于常规思维链的 Ours-CoT-7B 模型性能优于基线模型 Ours-ST-7B。而本文最终提出的 Ours-RobustCoT-7B 模型则在所有语言方向上均取得了最佳性能，相较于 Ours-CoT-7B 平均提升了 0.66 BLEU 分。这一结果表明，本文提出的鲁棒思维链训练策略所带来的性能增益并非仅限于 CoVoST 2 的特定数据分布，而是提升了模型通用的语音翻译能力。该方法在新的、未曾见过的数据分布上依然有效，证明了其良好的泛化性。

6 总结与未来研究

本文针对大语言模型在语音翻译任务中采用思维链策略时，易于对中间生成的源语言转录文本产生过度依赖，从而引发错误传播的问题，提出了一种鲁棒思维链方法。该方法在训练阶段，通过对语音输入和思维链文本进行随机掩码，并引入正则化约束，迫使模型综合利用原始语音信号和不完整的文本线索进行翻译，从而增强了模型面对不完美思维链时的鲁棒性。在 CoVoST 2 数据集六个主要翻译方向上的实验结果表明，本方法相较于不使用思维链的基线系统，平均 BLEU 分提升了 2.78 分，相较于标准思维链方法，也取得了 0.92 BLEU 分的性能增益。此外，本方法在不同参数规模的 Qwen2.5 系列模型上的验证结果，以及在 CoVoST 2 全部 21 个语向上的扩展实验，均证明了其有效性、通用性与良好的可扩展潜力。同时，该方法还能在一定程度上降低思维链中语音识别的错误率。

尽管本文所提出的方法取得了显著成效，但仍存在一定的局限性。首先，现有的鲁棒性训练策略虽然能够有效缓解模型对思维链的过度依赖，但仍无法完全避免在中间转录文本存在严重错误时对最终翻译质量的负面影响。其次，本文的实验验证主要在公开数据集 CoVoST 2 上进行，其数据规模与真实世界的应用场景相比仍然有限。展望未来，本文计划从以下两方面展开进一步研究：一方面，探索更为先进的模型设计与训练方法，旨在进一步提升模型对思维链中错误的识别与纠正能力，实现更高程度的鲁棒性；另一方面，计划将本方法扩展至更大规模的语音翻译数据集上进行验证，以期在更接近真实应用的环境下进一步挖掘其性能潜力。

致　谢　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] Sperber M, Neubig G, Niehues J, Waibel A. Neural lattice-to-sequence models for uncertain inputs//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark, 2017: 1380-1389

[2] Cheng Y, Tu Z, Meng F, Zhai J, Liu Y. Towards robust neural machine translation//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia, 2018: 1756-1766

[3] Sperber M, Neubig G, Pham N, Waibel A. Self-attentional models for lattice inputs//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy, 2019: 1185-1197

[4] Dong Q, Wang F, Yang Z, Chen W, Xu S, Xu B. Adapting translation models for transcript disfluency detection//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019: 6351-6358

[5] Zhang P, Ge N, Chen B, Fan K. Lattice transformer for speech translation//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy, 2019: 6475-6484

[6] Lam T, Schamoni S, Riezler S. Cascaded models with cyclic feedback for direct speech translation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada, 2021: 7508-7512

[7] Liu Yang. Recent Advances in Neural Machine Translation. Journal of Computer Research and Development, 2017, 54(6): 1144-1149 (in Chinese)
(刘洋. 神经机器翻译前沿进展. 计算机研究与发展, 2017, 54(6): 1144-1149)

[8] Li Ya-Chao, Xiong De-Yi, Zhang Min. A Survey of Neural Machine 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] Feng Yang, Shao Chen-Ze. Frontiers in Neural Machine Translation: A Literature Review. Journal of Chinese Information Processing, 2020, 34(7): 1-18 (in Chinese)
(冯洋, 邵晨泽. 神经机器翻译前沿综述. 中文信息学报, 2020, 34(7): 1-18)

[10] Ma Hou-Li, Dong Ling, Wang Jian, Wang Wen-Jun, Gao Sheng-Xiang, Yu Zheng-Tao. A Vietnamese-English End-to-end Speech Translation Method Based on Multi-feature Fusion. Journal of

Chinese Information Processing, 2024, 38(10): 35-45 (in Chinese)
(马候丽, 董凌, 王剑, 王文君, 高盛祥, 余正涛. 多特征融合的越英端到端语音翻译方法. 中文信息学报, 2024, 38(10): 35-45)

[11] Liu Yu-Chen, Zong Cheng-Qing. End-to-end Speech Translation by Integrating Cross-modal Information. Journal of Software, 2023, 34(4): 1837-1849 (in Chinese)
(刘宇宸, 宗成庆. 跨模态信息融合的端到端语音翻译. 软件学报, 2023, 34(4): 1837-1849)

[12] Li Ning, Zhu Li-Ping, Zhao Xiao-Bing, Renzeng Zhuo-Ma, Wang Yan-Min. End-to-end Speech Translation for Low-Resource Languages Based on Target Language Pre-Training and Joint Decoding. Journal of Chinese Information Processing, 2023, 37(12): 36-43 (in Chinese)
(李宁, 朱丽平, 赵小兵, 仁曾卓玛, 王燕敏. 基于目标语言预训练和联合解码的低资源语言端到端语音翻译. 中文信息学报, 2023, 37(12): 36-43)

[13] He Wen-Long, Gao Chang-Feng, Li Ta, Liu Jian. End-to-end Speech Translation Based on Adversarial Training. Journal of Signal Processing, 2021, 37(5): 893-901 (in Chinese)
(何文龙, 高长丰, 黎塔, 刘建. 基于对抗训练的端到端语音翻译研究. 信号处理, 2021, 37(5): 893-901)

[14] Wang C, Wu Y, Liu S, Yang Z, Zhou M. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020: 9161-9168

[15] Wang C, Wu Y, Liu S, Zhou M, Yang Z. Curriculum pre-training for end-to-end speech translation//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Online, 2020: 3728-3738

[16] Dong Q, Wang M, Zhou H, Xu S, Xu B, Li L. Consecutive decoding for speech-to-text translation//Proceedings of The Thirty-fifth AAAI Conference on Artificial Intelligence (AAAI). Online, 2021: 12738-12748

[17] Dong Q, Ye R, Wang M, Zhou H, Xu S, Xu B, Li L. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Online, 2021: 12749-12759

[18] Han C, Wang M, Ji H, Li L. Learning shared semantic space for speech-to-text translation//Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online, 2021: 2214-2225

[19] Inaguma H, Kawahara T, Watanabe S. Source and target bidirectional knowledge distillation for end-to-end speech translation//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). Online, 2021: 1872-1881

[20] Tang Y, Pino J, Li X, Wang C, Genzel D. Improving speech translation by understanding and learning from the auxiliary text translation task//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL). Online, 2021: 4252-4261

[21] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser Ł, Polosukhin I. Attention is all you need//Advances in Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017: 6000-6010

[22] Ye R, Wang M, Li L. End-to-end speech translation via cross-modal progressive training//Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech). Brno, Czechia, 2021: 2267-2271

[23] Xu C, Hu B, Li Y, Zhang Y, Huang S, Ju Q, Xiao T, Zhu J. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL). Online, 2021: 2619-2630

[24] Fang Q, Ye R, Li L, Feng Y, Wang M. Stemm: Self-learning with speech-text manifold mixup for speech translation//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland, 2022: 7050-7062

[25] Fang Q, Feng Y. Understanding and bridging the modality gap for speech translation//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 15864-15881

[26] Jiao W, Wang W, Huang J, Wang X, Shi S, Tu Z. Is chatgpt a good translator? Yes with gpt-4 as the engine. arXiv preprint arXiv:2301.08745, 2023.

[27] Wu J, Gaur Y, Chen Z, Zhou L, Zhu Y, Wang T, Li J, Liu S, Ren B, Liu L, et al. On decoder-only architecture for speech-to-text and large language model integration//Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Taipei, China, 2023: 1-8

[28] Wang M, Han W, Shafran I, Wu Z, Chiu C, Cao Y, Chen N, Zhang Y, Soltau H, Rubenstein P, et al. Slm: Bridge the thin gap between speech and text foundation models//Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Taipei, China, 2023: 1-8

[29] Chen Z, Huang H, Andrusenko A, Hrinchuk O, Puvvada K, Li J, Ghosh S, Balam J, Ginsburg B. Salm: Speech-augmented language model with in-context learning for speech recognition and translation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea, 2024: 13521-13525

[30] Gaido M, Papi S, Negri M, Bentivogli L. Speech translation with speech foundation models and large language models: What is there and what is missing?//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand, 2024: 14760-14778

[31] Radford A, Kim J, Xu T, Brockman G, McLeavey C, Sutskever I.

Robust speech recognition via large-scale weak supervision//Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA, 2023: 28492-28518

[32] Baevski A, Zhou Y, Mohamed A, Auli M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems (NeurIPS). Online, 2020: 12449-12460

[33] Tang C, Yu W, Sun G, Chen X, Tan T, Li W, Lu L, MA Z, Zhang C. SALMONN: Towards generic hearing abilities for large language models//Proceedings of The Twelfth International Conference on Learning Representations (ICLR). Vienna, Austria, 2024

[34] Du Y, Ma Z, Yang Y, Deng K, Chen X, Yang B, Xiang Y, Liu M, Qin B. Cot-st: Enhancing llm-based speech translation with multimodal chain-of-thought. arXiv preprint arXiv:2409.19510, 2024.

[35] Hu K, Chen Z, Yang C, Zelasko P, Hrinchuk O, Lavrukhin V, Balam J, Ginsburg B. Chain-of-thought prompting for speech translation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India, 2025: 1-5

[36] Berard A, Pietquin O, Servan C, Besacier L. Listen and translate: A proof of concept for end-to-end speech-to-text translation//Proceedings of the NIPS workshop on End-to-end Learning for Speech and Audio Processing. Barcelona, Spain, 2016

[37] Weiss R, Chorowski J, Jaitly N, Wu Y, Chen Z. Sequence-to-sequence models can directly translate foreign speech//Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech). Stockholm, Sweden, 2017: 2625-2629

[38] Deng P, Chen S, Zhang W, Zhang J, Dai L. The ustc's dialect speech translation system for iwslt 2023//Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). Toronto, Canada, 2023: 102-112

[39] Bansal S, Kamper H, Livescu K, Lopez A, Goldwater S. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Minneapolis, USA, 2019: 58-68

[40] Stoian M, Bansal S, Goldwater S. Analyzing asr pretraining for low-resource speech-to-text translation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020: 7909-7913

[41] Tang Y, Gong H, Dong N, Wang C, Hsu W, Gu J, Baevski A, Li X, Mohamed A, Auli M, Pino J. Unified speech-text pre-training for speech translation and recognition//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland, 2022: 1488-1499

[42] Le P, Gong H, Wang C, Pino J, Lecouteux B, Schwab D. Pre-training for speech translation: CTC meets optimal transport//Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA, 2023: 18667-18685

[43] Le H, Pino J, Wang C, Gu J, Schwab D, Besacier L. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation//Proceedings of the 28th International Conference on Computational Linguistics (COLING). Barcelona, Spain (Online), 2020: 3520-3533

[44] Zhang Y, Xu C, Li B, Chen H, Xiao T, Zhang C, Zhu J. Rethinking and improving multi-task learning for end-to-end speech translation//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, 2023: 10753-10765

[45] Liu Y, Xiong H, He Z, Zhang J, Wu H, Wang H, Zong C. End-to-end speech translation with knowledge distillation. arXiv preprint arXiv:1904.08075, 2019.

[46] Inaguma H, Kawahara T, Watanabe S. Source and target bidirectional knowledge distillation for end-to-end speech translation//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). Mexico City, Mexico, 2021: 1872-1881

[47] Lei Y, Xue Z, Zhao X, Sun H, Zhu S, Lin X, Xiong D. CKDST: Comprehensively and effectively distill knowledge from machine translation to end-to-end speech translation//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada, 2023: 3123-3137

[48] Jia Y, Johnson M, Macherey W, Weiss R, Cao Y, Chiu C, Ari N, Laurenzo S, Wu Y. Leveraging weakly supervised data to improve end-to-end speech-to-text translation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK, 2019: 7180-7184

[49] Bahar P, Zeyer A, Schluter R, Ney H. On using specaugment for end-to-end speech translation//Proceedings of the 16th International Conference on Spoken Language Translation (IWSLT). Hong Kong,China, 2019

[50] Lam T, Schamoni S, Riezler S. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland, 2022: 245-254

[51] Fang Q, Feng Y. Back translation for speech-to-text translation without transcripts//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 4567-4587

[52] Zhou Y, Fang Q, Feng Y. Cmot: Cross-modal mixup via optimal transport for speech translation//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 7873-7887

[53] Ye R, Wang M, Li L. Cross-modal contrastive learning for speech translation//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies (NAACL). Seattle, United States, 2022: 5099-5113

[54] Ouyang S, Ye R, Li L. Waco: Word-aligned contrastive learning for speech translation//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 3891-3907

[55] Bapna A, Chung Y, Wu N, Gulati A, Jia Y, Clark J, Johnson M, Riesa J, Conneau A, Zhang Y. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. arXiv preprint arXiv:2110.10329, 2021.

[56] Bapna A, Cherry C, Zhang Y, Jia Y, Johnson M, Cheng Y, Khanuja S, Riesa J, Conneau A. mslam: Massively multilingual joint pre-training for speech and text. arXiv preprint arXiv:2202.01374, 2022.

[57] Chen Z, Zhang Y, Rosenberg A, Ramabhadran B, Moreno P, Bapna A, Zen H. Maestro: matched speech text representations through modality matching//Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech). Incheon, Republic of Korea, 2022: 4093-4097

[58] Zhang Z, Zhou L, Ao J, Liu S, Dai L, Li J, Wei F. Speechut: bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Abu Dhabi, United Arab Emirates, 2022: 1663-1676

[59] Zhang Y, Kou K, Li B, Xu C, Zhang C, Xiao T, Zhu J. Soft alignment of modality space for end-to-end speech translation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea, 2024: 11041-11045

[60] Deng P, Zhang J, Zhou X, Ye Z, Zhang W, Cui J, Dai L. Learning Semantic Information from Machine Translation to Improve Speech-to-Text Translation//2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Taipei, China, 2023: 954-959

[61] Zhu Q S, Zhou L, Zhang Z, Liu S, Jiao B, Zhang J, Dai L, Jiang D, Li J, Wei F. Vatlm: visual-audio-text pre-training with unified masked prediction for speech representation learning. IEEE Transactions on Multimedia, 2024, 26: 1055-1064

[62] Zhang H, Si N, Chen Y, Zhang W, Yang X, Qu D, Jiao X. Tuning large language model for end-to-end speech translation. arXiv preprint arXiv:2310.02050, 2023.

[63] Huang Z, Ye R, Ko T, Dong Q, Cheng S, Wang M, Li H. Speech translation with large language models: An industrial practice. arXiv preprint arXiv:2312.13585, 2023.

[64] Chen X, Zhang S, Bai Q, Chen K, Nakamura S. Llast: Improved end-to-end speech translation system leveraged by large language models// Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand, 2024: 6976-6987

[65] Lam T, Gaido M, Papi S, Bentivogli L, Haddow B. Prepending or cross-attention for speech-to-text? an empirical comparison//Proceedings of the Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL). Albuquerque, New Mexico, 2025: 2994-3006

[66] Liu H, Chen A, Chen K, Bai X, Zhong M, Qiu Y, Zhang M. Adaptive inner speech-text alignment for llm-based speech translation. arXiv preprint arXiv:2503.10211, 2025.

[67] Zhang W, Naagar S, Ye Z, Tang P, Zhou X, Liu J, Dai L. Bridging modality gap with large speech and language models for end-to-end speech-to-text translation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India, 2025: 1-5

[68] Qwen. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[69] Wang C, Wu A, Pino J. Covost 2: A massively multilingual speech-to-text translation corpus//Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech). Brno, Czechia, 2021: 2247-2251

[70] Di Gangi M.A., Cattoni R., Bentivogli L., Negri M., Turchi M. 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 (NAACL). Minneapolis, Minnesota, 2019: 2012-2017

[71] Loshchilov I, Hutter F. Decoupled weight decay regularization//Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, USA, 2019

[72] Papineni K, Roukos S, Ward T, Zhu W. Bleu: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, Pennsylvania, USA, 2002: 311-318

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

[74] Chu Y, Xu J, Yang Q, Wei H, Wei X, Guo Z, Leng Y, Lv Y, He J, Lin J, Zhou C, Zhou J. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759, 2024.

[75] Barrault L, Chung Y, Cora Meglioli M, Dale D, Dong N, Duquenne P, Elsahar H, Gong H, Heffernan K, Hoffman J, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. arXiv preprint arXiv:2308.11596, 2023.

[76] Iranzo-Sánchez J., Silvestre-Cerdà J.A., Jorge J., Rosselló N., Giménez A., Sanchis A., Civera J., Juan A. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020: 8229-8233

FANG Qing-Kai, Ph.D. candidate. His research interests are natural language processing, speech translation and large language models.

FENG Yang, Ph.D., Professor, Ph.D. supervisor. Her research interests are natural language processing and large language models.

Background

The field of Speech Translation (ST), which aims to convert speech from a source language directly into text in a target language, is a critical area of research at the intersection of speech processing and natural language processing. Its importance is continually growing due to its vital applications in facilitating cross-lingual communication in scenarios such as international conferences, online education, and global business. Internationally, the field has progressed from traditional cascaded systems to end-to-end models. More recently, the advent of Large Language Models (LLMs) has ushered in a new era, with Chain-of-Thought (CoT) prompting emerging as a powerful technique. However, while CoT improves performance by breaking down the task, it introduces a critical vulnerability: the model's tendency to over-rely on the intermediate transcription, causing transcription errors to propagate and degrade the final translation quality and robustness.

This paper directly addresses this limitation by proposing a novel Robust Chain-of-Thought (RobustCoT) method for LLM-based speech translation. Our approach enhances the model's robustness by reducing its dependence on the intermediate transcription and forcing it to leverage the original acoustic information. This is achieved through a training strategy that combines random masking of the CoT with a Kullback-Leibler (KL) divergence regularization term, which ensures predictive consistency. Extensive experiments on the CoVoST 2 benchmark demonstrate that our method substantially improves translation quality, yielding an average BLEU score increase of 2.78 points over a standard baseline and 0.92 points over the conventional CoT method, with performance gains being most pronounced in high-transcription-error scenarios.

The implications of this research extend beyond the immediate task of ST. By developing a method to enhance the reliability of a sequential reasoning process, this work offers valuable insights for improving the robustness of other complex, multi-step generation tasks where the integrity of an intermediate chain of reasoning is crucial. The principles of using targeted masking and distributional consistency as regularization can inspire further innovations in building more trustworthy and reliable AI systems that are less susceptible to single-point-of-failure errors in their internal processing steps.

Furthermore, enhancing the robustness of speech translation technology has significant practical value. More reliable and accurate translation systems can significantly lower communication barriers in critical multilingual settings, from international diplomacy and business negotiations to emergency response and healthcare. This research, by making ST systems less fragile, contributes directly to the goal of fostering more seamless and effective global collaboration and understanding, ultimately impacting various domains that rely on clear cross-lingual communication.

Our research group has previously contributed to the fields of machine translation and speech translation, establishing a strong foundation for the current work. This study is part of our broader, ongoing effort to advance the state-of-the-art in cross-modal language technologies. The RobustCoT method presented here solves a key challenge within this larger agenda, pushing the boundaries of what is possible for robust, next-generation speech translation systems and offering a practical solution that can drive future research and real-world applications.