基于两种新颖辅助任务的端到端语音翻译

窦怀厦，吕孟哲，李军辉
苏州大学计算机科学与技术学院，江苏省苏州市
{20225227069,20225227080}@stu.suda.edu.cn，lijunhui@suda.edu.cn

摘要

端到端语音翻译具有跨模态和跨语言的特性，如何有效地利用这些特性是一个具有挑战性的问题。本文基于多任务学习框架，提出两种新颖辅助任务。语音增强的文本翻译任务通过在文本翻译任务中融入语音模态信息来缓解语音和文本的模态差异，最终提升语音翻译任务的性能。全局感知条件掩码语言建模任务能够同时建模转录文本和译文进而利用文本的全局上下文信息指导翻译模型的训练。在MuST-C数据集8个语向的实验结果表明，本文的方法显著优于基线系统，并且达到了与其它端到端语音翻译系统可竞争的性能水平。进一步的分析实验表明，本文的方法能够缓解语音和文本之间的模态差异并且在不损害文本翻译任务性能的情况下提升语音翻译任务的性能。

关键词： 端到端语音翻译；多任务学习；辅助任务

End-to-End Speech Translation Enhanced by Two Novel Auxiliary Tasks

Huaixia Dou, Mengzhe Lyu, Junhui Li
School of Computer Science and Technology, Soochow University, Suzhou, China
{20225227069,20225227080}@stu.suda.edu.cn，lijunhui@suda.edu.cn

Abstract

End-to-end speech translation is characterized by cross-modal and cross-linguistic properties. How to better utilize these properties is a challenging question. In this paper, we propose two novel auxiliary tasks based on a multi-task learning framework. The speech augmented text translation task alleviates the modality gap between speech and text by incorporating speech modality information into the text translation task, which ultimately improves the performance of the speech translation task. The globally aware conditional mask language modeling task is able to model both the transcription and the translation simultaneously and then use the global context information of the text to guide the training of the translation model. Experimental results on the MuST-C benchmark show that the approach in this paper significantly outperforms the baseline system and achieves a competable performance with other end-to-end speech translation systems. Further analysis shows that our approach is able to alleviate the modality gap between speech and text and preserve the performance of the original machine translation task.

Keywords: End-to-end speech translation , Multi-task learning , Auxiliary tasks

第二十三届中国计算语言学大会论文集，第183页-第196页，太原，中国，2024年7月25日至28日。
卷1：主会论文

1 引言

语音翻译旨在将源语言的语音转化为目标语言的文本。传统的语音翻译系统通常采用级联方法 (Dong et al., 2019; Zhang et al., 2019; Sperber and Paulik, 2020; Lam et al., 2021)，首先使用自动语音识别系统将源语言的语音转录为文本，接着使用机器翻译系统将转录文本翻译为目标语言的文本。由于包含多个中间步骤，级联系统存在错误传播和高延迟的问题。端到端语音翻译因为无需生成中间的转录文本，有望克服级联系统的缺陷，因此在近年来受到越来越多研究者的关注 (Wang et al., 2020b; Ren et al., 2020; Zhang et al., 2020a; Huang et al., 2021; Lam et al., 2022; Han et al., 2023)。目前端到端语音翻译的主流做法是在多任务学习框架下引入多个相关任务来改善语音翻译任务的训练，本文的目标是提出更好的辅助任务来提升语音翻译任务的性能。

在多任务学习框架下，语音翻译任务是主任务，而文本翻译任务通常被用来辅助语音翻译任务的训练 (Ye et al., 2021; Tang et al., 2021; Ye et al., 2022; Yin et al., 2023; Zhang et al., 2023b)。传统的文本翻译任务只涉及单一模态，而语音模态中富含的语义信息可以为文本翻译任务提供额外的上下文和语音特征。因此，本文提出语音增强的文本翻译(Speech Augmented Text Translation, SATT)任务，在文本翻译过程中融合语音模态的信息来缓解语音和文本的模态差异，进而提升语音翻译任务的性能。此外，当前的工作主要集中在同时对跨模态的源端语音和转录文本进行建模，而对同模态的转录文本和译文关注较少。Zhou等人(2022)研究表明，译文中包含的上下文信息能够进一步提升文本翻译任务的性能。因此，本文提出全局感知条件掩码语言建模(Globally Aware Conditional Masked Language Modeling, GACMLM)任务来同时建模转录本文和译文并且利用文本的全局上下文信息来辅助翻译模型的训练。

具体地，针对SATT任务，本文对编码器中的自注意力模块进行修改，使得在自注意力计算过程中，转录文本能够同时关注到文本模态和语音模态的源端上下文信息。针对GACMLM任务，本文借鉴Zhou等人(2022)的思路，采用条件掩码语言模型 (Ghazvininejad et al., 2019)来建模转录文本和译文的双向全局上下文。整个训练过程分为三个阶段。在第一个阶段，利用外部机器翻译语料来预训练模型的翻译编码器和翻译解码器模块；在第二个阶段，利用语音增强的文本翻译任务辅助语音翻译任务的训练，此外在该阶段以多任务学习的方式同时训练条件掩码语言模型；在第三个阶段，将训练好的条件掩码语言模型作为教师模型，通过知识蒸馏 (Hinton et al., 2015)同时辅助语音翻译任务和语音增强的文本翻译任务的训练。在MuST-C八个语向的实验结果表明，本文方法的平均BLEU值能够达到30.4，相比基线系统提升1.5个点。进一步的分析实验表明，本文的方法能够缓解语音和文本的模态差异，同时在不损害文本翻译任务性能的情况下提升语音翻译任务的性能。

2 相关工作

2.1 端到端语音翻译

为了应对级联系统中固有的错误传播和高时延等问题，端到端语音翻译被提出来直接将源语言的语音映射为目标端文本，而无需生成中间的临时转录文本。作为跨模态和跨语言的任务，端到端语音翻译可获得的数据远少于语音识别和文本翻译的数据。为了应对语音翻译数据不足的问题，Ye等人 (2019)；Bahar等人 (2019)；Pino等人(2019)利用语音识别和文本翻译数据合成了更多的语音翻译数据，Pino等人 (2020)；Wang等人(2021)使用自训练技术进行数据增强，Lam等人 (2022)采用反向翻译和知识蒸馏来获取额外的训练数据。为了应对建模困难问题 (Xu et al., 2023)，多任务学习框架被广泛使用 (Ye et al., 2021; Ye et al., 2022; Fang et al., 2022; Zhou et al., 2023)。在多任务框架下，文本翻译任务通常作为辅助任务，在此基础上，为了缩小语音和文本的模态差异，Han等人 (2021)使用共享的语义记忆模块把语音和文本特征投影到公共的语义空间中，Ye等人 (2022)；Quyang等人 (2023)；Zhang等人 (2022a)采用对比学习方法来拉近语音表示和文本表示的距离，Fang等人 (2022)；Zhou等人 (2023)；Cheng等人 (2023)采用混合语音序列和文本序列的方法来缓解跨模式表征差异。区别于先前的工作，本文在常规的文本翻译任务中利用注意力机制同语音模态进行交互。一方面语音模态中富含的语义信息可以为文本翻译任务提供额外的上下文和语音特征，另一方面转录文本与语音模态的交

项目资助：江苏高校优势学科建设工程资助项目

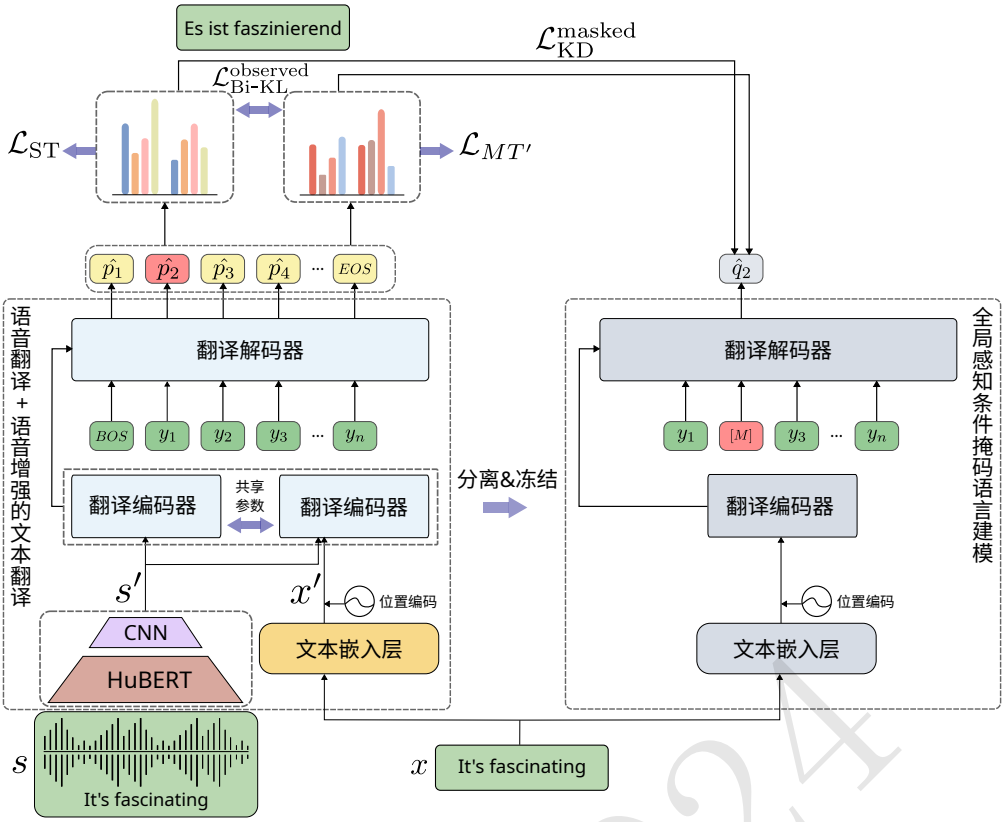

图 1. 本文模型架构(第三阶段)

互能够缩小语音和文本的模态差异。此外，相关工作中很少涉及到建模目标端译文，而译文中包含的上下文信息能够为翻译任务提供额外的帮助。

2.2 条件掩码语言模型

条件掩码语言模型通常采用传统的编码器-解码器架构，利用源端输入文本作为条件，目标端可见单词作为上下文来预测被遮盖的单词。相比于神经翻译模型，条件掩码语言模型不仅能够关注源端的文本信息，而且能够关注目标端的上下文信息。Ghazvininejad等人 (2019)使用条件掩码语言模型预测被遮盖的单词，并结合高效的并行迭代解码算法来非自回归地生成译文。Higuchi等人 (2020)组合连接时序分类(CTC)和条件掩码语言模型来改善CTC的预测结果。Chen等人 (2021)将其用于非自回归语音识别任务并且额外增加约束来缩小训练与推理时的差异。Inaguma等人 (2021)将其用于端到端语音翻译的非自回归解码。Chen等人 (2020)；Zhou等人(2022)将其作为教师模型，利用知识蒸馏来改善文本生成任务和文本翻译任务的训练。本文受Zhou等人 (2022)启发，提出全局感知条件掩码语言建模任务来建模文本的全局上下文信息，并且利用这部分信息来辅助翻译模型的训练。

3 方法

本文在多任务学习框架下提出了两个新颖辅助任务，即语音增强的文本翻译任务和全局感知条件掩码语言建模任务，旨在改善语音翻译任务的训练。本节首先对端到端语音翻译进行建模，接着依次介绍这几个辅助任务。

3.1 问题建模

语音翻译的训练数据通常由包含源语言语音、源语言转录文本和目标语言译文的三元组构成，形式化表示为$\mathcal{D}_{\text{ST}} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$。其中$\mathbf{s} = (s_1, s_2, ..., s_{|\mathbf{s}|})$是语音帧序列，$\mathbf{x} = (x_1, x_2, ..., x_{|\mathbf{x}|})$是转录文本序列，$\mathbf{y} = (y_1, y_2, ..., y_{|\mathbf{y}|})$是译文序列，$|\mathbf{s}|$、$|\mathbf{x}|$和$|\mathbf{y}|$分别是语音帧序列、转录文本序列和译文序列的长度。端到端语音翻译系统需要直接将语音帧序列$\mathbf{s}$转换为译

文序列$\mathbf{y}$而无需生成中间的转录文本序列$\mathbf{x}$。除此之外，本文还利用外部机器翻译数据对模型进行预训练，这部分数据记为$\mathcal{D}_{\mathrm{EXT}}=\{(\mathbf{x}^{\mathbf{ext}},\mathbf{y}^{\mathbf{ext}})\}$。

3.2 语音翻译

本文使用声学编码器来提取语音特征，利用翻译编码器对语音特征进一步编码，最终通过翻译解码器来完成目标译文的解码。

3.2.1 编码过程

预训练的HuBERT (Hsu et al., 2021)模型和两个额外的卷积层作为声学编码器来提取语音$\mathbf{s}$的底层特征。原始波形信号的采样频率为16kHz，经过HuBERT模型后得到语音特征序列。由于语音特征序列的长度通常远大于其转录文本的长度，为了进一步匹配语音特征序列和转录文本的长度，本文在HuBERT之后使用两个卷积层来缩短语音特征序列的长度。声学编码器和卷积层对语音$\mathbf{s}$的编码过程如公式(1)所示：

$$\mathbf{s}'=\mathrm{CNN}\left(\mathrm{HuBERT}\left(\mathbf{s}\right)\right), \tag{1}$$

其中，HuBERT(·)表示使用HuBERT模型进行编码，CNN(·)表示卷积操作。

翻译编码器用于进一步对声学编码器的输出$\mathbf{s}'$编码，翻译编码器在结构上与标准的Transformer (Vaswani et al., 2017)编码器相同，由N_e个完全相同的层堆叠而成，每层由一个多头掩码自注意力子层和一个全连接前馈神经网络构成。在语音翻译任务中，多头掩码自注意力子层使用前一层的隐状态作为输入并通过多头注意力模块和层规范化模块进行计算，如公式(2)所示：

$$B^{(i)}=\mathrm{LayerNorm}\left(\mathrm{MultiHead}\left(A^{(i-1)},A^{(i-1)},A^{(i-1)}\right)+A^{(i-1)}\right), \tag{2}$$

其中，$1<=i<=N_e$，$A^{(i-1)}\in\mathbb{R}^{t_s\times d_{\mathrm{model}}}$表示第$i-1$层的隐状态，$t_s$为经过卷积层后的语音特征序列的长度，$d_{\mathrm{model}}$为模型维度，$A^{(0)}=\mathbf{s}'$为声学编码器的输出，MultiHead(·)表示多头注意力模块，LayerNorm(·)表示层规范化，$B^{(i)}\in\mathbb{R}^{t_s\times d_{\mathrm{model}}}$。接下来，全连接前馈神经网络使用$B^{(i)}$作为输入并经过层规范化模块计算得到当前层的语音编码输出$A^{(i)}$，如公式(3)所示：

$$A^{(i)}=\mathrm{LayerNorm}\left(\mathrm{FeedForward}\left(B^{(i)}\right)+B^{(i)}\right), \tag{3}$$

其中，LayerNorm(·)表示层规范化，FeedForward(·)表示全连接前馈神经网络。为了简便起见，本文定义$\mathbf{a}=A^{(N_e)}$表示语音经过翻译编码器后得到的上下文表示。

3.2.2 解码过程

翻译解码器使用标准的Transformer解码器，由N_d个完全相同的层堆叠而成。每层由一个多头掩码自注意力子层，一个多头上下文注意力子层和一个全连接前馈神经网络构成。本文使用交叉熵损失作为语音翻译任务的损失函数，如公式(4)所示：

$$\mathcal{L}_{\mathrm{ST}}(\mathbf{s},\mathbf{y})=-\sum_{i=1}^{|\mathbf{y}|}\log P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i}). \tag{4}$$

3.3 语音增强的文本翻译

为了缓解模态差异，本文提出语音增强的文本翻译任务来融入语音模态的信息。在多任务学习框架下，语音增强的文本翻译任务和语音翻译任务共享翻译编码器与翻译解码器。

3.3.1 编码过程

文本词嵌入层和位置编码模块用来编码转录文本$\mathbf{x}$，如公式(5)所示：

$$\mathbf{x}'=\mathrm{Embedding}\left(\mathbf{x}\right)+\mathrm{Pos}\left(\mathbf{x}\right), \tag{5}$$

其中，Embedding(·)表示文本词嵌入层，Pos(·)表示位置编码模块。

为了有效地利用语音模态中富含的语义信息，本文在对转录文本进行多头自注意力计算时，通过将语音和转录文本的键值对向量进行拼接来融合语音模态的信息。具体地，声学编码器的输出$\mathbf{s}'$将和翻译编码器中每一层的隐状态输出$A^{(i-1)}$进行拼接作为新的文本键值对向量参与多头注意力计算，如公式(6)和公式(7)所示：

$$Q^{(i-1)} = A^{(i-1)}, K^{(i-1)} = V^{(i-1)} = \text{Concat}(\mathbf{s}', A^{(i-1)}), \tag{6}$$

$$B^{(i)} = \text{LayerNorm}(\text{MultiHead}(Q^{(i-1)}, K^{(i-1)}, V^{(i-1)}) + A^{(i-1)}), \tag{7}$$

其中，$1 <= i <= N_e$，$A^{(0)} = \mathbf{x}'$。接下来利用公式(3)，以$B^{(i)}$作为输入并经过层规范化模块计算得到当前层的文本编码输出$A^{(i)}$。为了简便起见，本文定义$\mathbf{e_{mix}} = A^{(N_e)}$表示文本经过翻译编码器并且进行语音增强后得到的上下文表示。

3.3.2 解码过程与模态正则化

翻译解码器以$\mathbf{e_{mix}}$作为输入，通过交叉熵损失函数来计算语音增强的文本翻译任务的损失，如公式(8)所示：

$$\mathcal{L}_{\text{MT}'}(\mathbf{s}, \mathbf{x}, \mathbf{y}) = -\sum_{i=1}^{|\mathbf{y}|} \log P(\mathbf{y}_i | \mathbf{e_{mix}}, \mathbf{y}_{<i}). \tag{8}$$

为了缓解语音和文本的模态差异，本文希望模型分别以语音上下文表示$\mathbf{a}$和以语音增强后的文本上下文表示$\mathbf{e_{mix}}$作为条件进行预测的目标端概率分布趋于一致。本文使用双向KL散度作为模态正则化损失来衡量两个预测概率的不一致性程度，如公式(9)所示：

$$\begin{aligned}\mathcal{L}^{\text{all}}_{\text{Bi-KL}}(\mathbf{s}, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{y}|} \frac{1}{2} (&\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i | \mathbf{a}, \mathbf{y}_{<i}) || P(\mathbf{y}_i | \mathbf{e_{mix}}, \mathbf{y}_{<i})) \\ &+ \mathcal{D}_{\text{KL}}(P(\mathbf{y}_i | \mathbf{e_{mix}}, \mathbf{y}_{<i}) || P(\mathbf{y}_i | \mathbf{a}.\mathbf{y}_{<i}))).\end{aligned} \tag{9}$$

3.4 全局感知条件掩码语言建模

为了从文本模态中学习更多的上下文信息，本文提出全局感知条件掩码语言建模任务来同时建模源端的转录文本和目标端的译文。该任务可分为文本全局上下文学习与知识蒸馏两个子任务。文本全局上下文学习任务通过训练条件掩码语言模型来学习文本的全局上下文信息。知识蒸馏任务通过分离出训练好的条件掩码语言模型并使用该模型进一步指导语音翻译任务和语音增强的文本翻译任务的训练。

3.4.1 文本全局上下文学习

条件掩码语言模型按照概率p随机对目标端译文执行掩码操作。目标端译文$\mathbf{y}$中的单词将被划分为两个集合。其中，掩码单词集合记为$\mathbf{y}_{masked}$，可见单词集合记为$\mathbf{y}_{observed}$。针对掩码单词集合$\mathbf{y}_{masked}$，其在原句中的对应位置被替换为特殊的掩码标记$[M]$。条件掩码语言模型会根据转录文本$\mathbf{x}$和目标端可见单词集合$\mathbf{y}_{observed}$来预测目标端掩码单词集合$\mathbf{y}_{masked}$。

与Zhou等人(2022)不同，本文的条件掩码语言模型和语音翻译任务共享文本词嵌入层、翻译编码器与翻译解码器。条件掩码语言模型首先使用文本词嵌入层和位置编码模块来对转录文本$\mathbf{x}$进行编码得到$\mathbf{x}'$，接着翻译编码器对$\mathbf{x}'$进一步编码得到文本的上下文表示$\mathbf{e}$。为了使翻译解码器适应条件掩码语言建模任务，本文移除了翻译解码器中多头掩码自注意力子层的注意力掩码，因此条件掩码语言模型能够同时关注到转录文本、译文的历史信息和译文的未来信息，它应该包含双向的文本全局上下文信息。本文使用交叉熵损失作为全局感知条件掩码语言建模任务的损失函数，如公式(10)所示：

$$\mathcal{L}_{\text{CMLM}}(\mathbf{x}, \mathbf{y}) = -\sum_{\mathbf{y}_i \in \mathbf{y}_{masked}} \log P(\mathbf{y}_i | \mathbf{e}, \mathbf{y}_{observed}). \tag{10}$$

3.4.2 知识蒸馏

在文本全局上下文学习完成后，本文从文本词嵌入层、翻译编码器和翻译解码器中分离出训练好的条件掩码语言模型。如图1右侧所示，分离出的条件掩码语言模型的参数不再更新，此时把条件掩码语言作为教师模型，利用知识蒸馏来进一步辅助左侧翻译模型的训练。针对掩码单词集合$\mathbf{y}_{masked}$，教师模型利用学习到的文本全局上下文信息来进一步改善语音翻译任务和语音增强的文本翻译任务的训练。如公式(11)所示：

$$\begin{aligned}\mathcal{L}_{\text{KD}}^{\text{masked}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \sum_{\mathbf{y}_i \in \mathbf{y}_{masked}} & (\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{e},\mathbf{y}_{observed})||P(\mathbf{y}_i|\mathbf{e}_{\mathbf{mix}},\mathbf{y}_{<i})) \\ & +\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{e},\mathbf{y}_{observed})||P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i}))).\end{aligned} \tag{11}$$

3.5 训练方式

整个训练过程分为三个阶段。在第一个阶段，采用外部机器翻译语料对模型的翻译编码器和翻译解码器进行训练。该阶段的损失函数如公式(12)所示：

$$\mathcal{L}_{\text{stage1}}(\mathbf{x}^{\mathbf{ext}},\mathbf{y}^{\mathbf{ext}}) = -\sum_{i=1}^{|\mathbf{y}^{\mathbf{ext}}|} \log P(\mathbf{y}_i^{\mathbf{ext}}|\mathbf{x}^{\mathbf{ext}},\mathbf{y}_{<i}^{\mathbf{ext}}). \tag{12}$$

在第二个阶段，通过多任务学习来联合训练语音翻译任务、语音增强的文本翻译任务和文本全局上下文学习任务。该阶段的损失函数如公式(13)所示：

$$\mathcal{L}_{\text{stage2}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \mathcal{L}_{\text{ST}}(\mathbf{s},\mathbf{y}) + \mathcal{L}_{\text{MT}'}(\mathbf{s},\mathbf{x},\mathbf{y}) + \mathcal{L}_{\text{Bi-KL}}^{\text{all}}(\mathbf{s},\mathbf{x},\mathbf{y}) + \mathcal{L}_{\text{CMLM}}(\mathbf{x},\mathbf{y}). \tag{13}$$

在第三个阶段，对于可见单词集合$\mathbf{y}_{observed}$，继续使用双向KL散度约束语音翻译任务和语音增强的文本翻译任务的目标端预测的概率分布。如公式(14)所示：

$$\begin{aligned}\mathcal{L}_{\text{Bi-KL}}^{\text{observed}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \sum_{\mathbf{y}_i \in \mathbf{y}_{observed}} & \frac{1}{2}(\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i})||P(\mathbf{y}_i|\mathbf{e}_{\mathbf{mix}},\mathbf{y}_{<i})) \\ & +\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{e}_{\mathbf{mix}},\mathbf{y}_{<i})||P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i}))),\end{aligned} \tag{14}$$

对于掩码单词集合$\mathbf{y}_{masked}$，利用知识蒸馏把训练好的条件掩码语言模型作为教师模型，进一步提升语音翻译任务和语音增强的文本翻译任务的性能。该阶段的损失函数如公式(15)所示：

$$\mathcal{L}_{\text{stage3}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \mathcal{L}_{\text{ST}}(\mathbf{s},\mathbf{y}) + \mathcal{L}_{\text{MT}'}(\mathbf{s},\mathbf{x},\mathbf{y}) + \mathcal{L}_{\text{Bi-KL}}^{\text{observed}}(\mathbf{s},\mathbf{x},\mathbf{y}) + \alpha\mathcal{L}_{\text{KD}}^{\text{masked}}(\mathbf{s},\mathbf{x},\mathbf{y}), \tag{15}$$

其中，α是超参数，控制知识蒸馏任务的权重[0]。

4 实验

4.1 数据集

语音翻译数据集 MuST-C (Di Gangi et al., 2019)是一个多语言的语音翻译数据集。参考相关的研究工作，本文在英语(EN)到其它八种语言的翻译方向上进行实验，这八种语言包括德语(DE)、法语(FR)、俄语(RU)、西班牙语(ES)、意大利语(IT)、罗马尼亚语(RO)、葡萄牙语(PT)和荷兰语(NL)。每个翻译方向都包含至少385个小时的录音数据。本文使用验证集对模型进行验证，并且在测试集tst-COMMON上报告最终结果。

机器翻译数据集 本文使用的外部机器翻译数据集包括用于英语-德语/法语/俄语/西班牙语/罗马尼亚语方向的WMT数据集和用于英语-意大利语/葡萄牙语/荷兰语的OPUS100数据集[1] (Zhang et al., 2020b)。详细的统计数据可参见表1，其中列出了语音翻译数据集和外部机器翻译数据集的相关信息。

预处理 本文使用原始格式的16位16kHz单声道音频作为语音输入。转录文本和译文均区分大小写。对于每个翻译方向，本文利用单字元(Unigram)SentencePiece[2]模型在语音翻译数据集上进行训练来学习转录文本和译文间的共享词表。接下来，使用共享词表把语音翻译语料和外部机器翻译语料分割成子词单元。本文将共享词表大小统一设置为10K。

[0]考虑到第三阶段$\mathcal{L}_{\text{Bi-KL}}^{\text{observed}}$与第二阶段$\mathcal{L}_{\text{Bi-KL}}^{\text{all}}$的区别仅仅在于译文$\mathbf{y}$中选中的单词范围不同，为了保持一致性，本文没有将其包含在α的控制范围

[1]http://opus.nlpl.eu/opus-100.php

[2]https://github.com/google/sentencepiece

英语 (EN)→	MuST-C		外部机器翻译数据	
	小时数	句子数	来源	句子数
德语 (DE)	408	234K	WMT16	4.6M
法语 (FR)	492	280K	WMT16	40.8M
俄语 (RU)	489	270K	WMT16	2.5M
荷兰语 (NL)	442	253K	OPUS100	1.0M
西班牙语 (ES)	504	270K	WMT13	15.2M
意大利语 (IT)	465	258K	OPUS100	1.0M
葡萄牙语 (PT)	385	211K	OPUS100	1.0M
罗马尼亚语 (RO)	432	240K	WMT16	0.6M

表 1. 实验数据集统计

4.2 实验设置

模型设置 本文使用预训练的HuBERT模型来提取语音特征。在HuBERT[3]之后添加了两层卷积层来进一步缩短语音特征序列的长度，卷积层内核大小为5，步长为2，填充大小为2，隐层维度为512。本文翻译编码器和翻译解码器层数N_e和N_d均设置为6，多头注意力设置为8个头，隐层维度设置为512、全连接前馈神经网络中间层维度设置为2048。

训练与推理 本文使用Adam优化器 (Kingma and Ba, 2015)，其中，$\beta_1 = 0.9, \beta_2 = 0.98$。在整个训练过程中，文本批处理大小设置为33K个令牌以内，语音批处理大小设置为16M语音帧以内，预热(WarmUp)步数设置为4K，Dropout和标签平滑(Label Smoothing)设置为0.1。第一阶段的学习率设置为7e-4，后续阶段的学习率设置为1e-4。为了避免模型过拟合，在第一阶段和第三阶段，当验证集上的BLEU值超过10轮不再提升则停止该阶段训练；在第二阶段，当验证集上的条件掩码语言模型损失值$\mathcal{L}_{\text{CMLM}}$超过5轮不再降低则停止该阶段训练。本文中知识蒸馏权重α和掩码概率p最终分别设置为0.5和0.15。与训练过程不同，推理过程中不再使用转录文本和译文信息，因此本文仅保留语音翻译模块进行推理。为了公平比较，本文平均最后10个检查点用于评估模型性能。本文使用集束搜索(Beam Search)并且设置Beam Size为8。针对八个翻译方向，本文设置了不同的长度惩罚值，分别是1.1(英语-德语)、1.7(英语-法语)、0.4(英语-俄语)、0.5(英语-西班牙语)、1.6(英语-罗马尼亚语)、0.7(英语-意大利语)、1.4(英语-葡萄牙语)、0.9(英语-荷兰语)。本文的模型使用FairSeq[4] (Ott et al., 2019)实现，并且在4张Nvidia Tesla V100上训练得到。

模型评估 对于所有翻译方向，本文使用SacreBLEU[5] (Post, 2018)来报告大小写敏感并且去标记化(Detokenized)后的BLEU分数 (Papineni et al., 2002)。

4.3 实验结果

为了公平起见，本文方法将同多个端到端语音翻译模型进行比较，这些模型均使用了外部机器翻译数据。我们根据模型使用的声学编码器分为两类：基于Wav2vec2.0 (Baevski et al., 2020)的模型和基于HuBERT的模型。

- 基于Wav2Vec2.0的模型：Chimera (Han et al., 2021)通过将语音和文本表征投影到共同的语义表征空间中来缩小模态差异；XSTNet (Ye et al., 2021)提出了一个多任务训练框架，并采用渐进式训练算法来提高语音翻译任务的性能；STEMM (Fang et al., 2022)通过混合语音和转录文本序列，并应用正则化方法来学习更好的语音表示；ConST (Ye et al., 2022)利用对比学习来缩小模态差异；FCGCL (Zhang et al., 2022a)利用多粒度对比学习来缩小模态差异；SpeechUT (Zhang et al., 2022b)利用有监督和无监督数据进行联合预训练，并通过统一的编码器来处理语音和文本模态。

[3]https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt
[4]https://github.com/pytorch/fairseq
[5]https://github.com/mjpost/sacrebleu, signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0

模型	BLEU								
	EN-DE	EN-FR	EN-RU	EN-ES	EN-IT	EN-RO	EN-PT	EN-NL	Avg.
基于Wav2Vec2.0									
Chimera (Han et al., 2021)	27.1	35.6	17.4	30.6	25.0	24.0	30.2	29.2	27.4
XSTNet (Ye et al., 2021)	27.1	38.0	18.5	30.8	26.4	25.7	32.4	31.2	28.8
STEMM (Fang et al., 2022)	28.7	37.4	17.8	31.0	25.8	24.5	31.7	30.5	28.4
ConST (Ye et al., 2022)	28.3	38.3	18.9	32.0	27.2	25.6	33.1	31.7	29.4
FCGCL (Zhang et al., 2022a)	28.7	37.5	19.1	31.2	26.5	26.0	32.1	31.0	29.0
SpeechUT (Zhang et al., 2022b)	**30.1**	**41.4**	-	**33.6**	-	-	-	-	-
W2V2-ST	27.1	37.7	18.5	31.8	26.2	24.4	30.7	30.4	28.4
本文方法	28.8*	39.2*	**19.9***	33.0*	**27.3***	**26.4***	**33.2***	**32.2***	**30.0**
基于HuBERT									
CMOT (Zhou et al., 2023)	29.0	39.5	19.2	32.8	27.5	26.0	33.5	32.1	30.0
CRESS (Fang and Feng, 2023)	29.4	40.1	19.7	33.2	27.6	**26.4**	33.6	32.3	30.3
DUB (Zhang et al., 2023a)	26.2	35.3	-	30.4	-	-	-	-	-
IMTL-KD (Zhang et al., 2023c)	**29.7**	**41.1**	-	**33.9**	-	-	-	-	-
HuBERT-ST	27.5	38.1	19.0	32.3	26.5	25.1	31.8	30.8	28.9
本文方法	29.2*	40.0*	**20.1***	33.3*	**27.7***	**26.4***	**33.8***	**32.4***	**30.4**

表 2. 在MuST-C测试集tst-COMMON上与其它方法性能比较，*表示较基线模型(即W2V2-ST或HuBERT-ST)的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

#	模型	BLEU
1	ST	25.0
2	CRESS	**26.7**
3	本文方法	26.6*

表 3. 在CoVoST 2英语-德语测试集上性能对比，*表示较#1的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

- 基于HuBERT的模型：CMOT (Zhou et al., 2023)在STEMM的基础上，利用最优传输理论自动找到语音和转录文本之间的对齐，并利用对齐信息来混合语音和转录文本序列；CRESS (Fang and Feng, 2023)把模态差异和机器翻译任务的暴露偏差问题关联起来，提出目标端计划采样、正则化方法、令牌级自适应训练来缓解模态差异。DUB (Zhang et al., 2023a)使用离散单元代替连续的语音序列，并通过反向翻译技术提高语音翻译任务的性能；IMTL-KD (Zhang et al., 2023c)分析了多任务训练方法中不同任务之间的一致性，为语音翻译任务提出了改进的多任务训练方法，该方法通过减少语音和文本的长度和表征差异来缩小模态差异。

参考Fang等人 (2022)的做法，本文使用相同的模型结构实现了类似的基线模型HuBERT-ST，HuBERT-ST仅仅对语音翻译任务进行模型参数优化。为了同基于Wav2vec2.0的模型进行比较，本文基于Wav2vec2.0[6]实现了本文的方法和基线系统W2V2-ST。实验结果如表2所示，从表2我们可以看出，

- 无论是基于Wav2Vec2.0还是基于HuBERT，本文模型明显优于相应的基线模型。相比于基线模型，本文模型在八个语向的BLEU平均分别提升1.6个点和1.5个点。
- 相比于其它端到端语音翻译模型，本文模型平均而言取得了最好的性能[7]。

5 分析与讨论

5.1 CoVoST 2英语-德语方向性能

参照CRESS (Fang and Feng, 2023)做法，本文除了在MuST-C数据集的八个语向上进行实验之外，还在更大的数据集CoVoST 2 (Wang et al., 2020a)上测试本文方法的性能。CoVoST

[6]https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt
[7]由于SpeechUT和IMTL-KD大部分语向性能并未给出，此处不考虑SpeechUT和IMTL-KD

#	模型	BLEU
1	本文方法	**29.2***
2	−GACMLM	28.9*
3	−SATT	28.5*
4	−GACMLM−SATT	27.5

表 4. 在MuST-C英语-德语测试集tst-COMMON上的消融实验结果，*表示较#4的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

#	模型	BLEU
1	ST	27.5
2	ST+TT	28.6*
3	ST+SATT	**28.9***

表 5. 在MuST-C英语-德语测试集tst-COMMON上性能对比，*表示较#1的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

2是一个大型多语种语音翻译数据集，涵盖从21种语音翻译成英语和从英语语音翻译成15种语言的内容。本文在英语-德语方向进行实验，该方向包含430小时的语音，并带有对应的转录文本和译文。本文使用开发集进行验证，使用测试集进行评估，使用与MuST-C相同的预处理、模型配置和超参数(见4.1)。结果如表3所示，表中"ST"表示基线模型，其仅在语音翻译任务上训练。从表中可以看出，本文方法的性能明显优于基线系统(+1.6)并且与CRESS相比具有可竞争的性能，表明本文的方法在不同数据集上的有效性。

5.2 消融分析

为了探究不同部分对总体模型的贡献，本文在英语-德语方向测试集tst-COMMON上进行消融实验。实验结果如表4所示。从表中可以看出，在移除全局感知条件掩码语言建模(GACMLM)任务后，BLEU下降了0.3个点，表明文本的全局上下文信息有利于语音翻译任务。在移除语音增强的文本翻译(SATT)任务后，BLEU下降了0.7个点，表明该任务对于总体模型影响较大。在移除本文提出的两个辅助任务后，BLEU急剧下降至27.5(-1.7)，表明在多任务学习框架下，联合使用本文提出的两个新颖辅助任务能够有效提升语音翻译任务的性能。

5.3 文本翻译中语音的作用

为了探究语音增强的文本翻译任务中语音起到的效果。本文在MuST-C测试集tst-COMMON上做了进一步的实验，实验结果如表5所示，表中各个模型均已经过第一阶段外部机器翻译语料预训练，具体代表的含义如下所示:

- ST：表示模型仅在语音翻译任务上训练(即HuBERT-ST)。
- ST+TT：表示模型在语音翻译任务、文本翻译任务和模态正则化损失上进行联合训练。
- ST+SATT：表示模型在语音翻译任务和语音增强的文本翻译任务上进行联合训练。

比较表中#1和#2，我们可以看出，在移除文本翻译任务后，BLEU下降了1.1，表明在多任务学习框架下文本翻译任务的加入能够有效提升语音翻译任务的性能。比较表中#2和#3，我们可以看出，在源端对转录文本进行自注意力计算时融入语音模态信息有利于语音翻译任务的性能提升(+0.3)，验证了本文提出的语音增强的文本翻译任务的有效性。

5.4 知识蒸馏权重α和掩码概率p对语音翻译性能的影响

本文提出全局感知条件掩码语言建模任务来改善语音翻译任务的训练。为了探究公式(15)中知识蒸馏权重α对语音翻译性能的影响，本文在MuST-C英语-德语验证集上进行网格搜索，令$\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$，实验结果如图2(左)所示。从图中可以看出，当$\alpha = 0.5$时，MuST-C验证集上性能最佳。为了探究条件掩码语言模型中单词掩

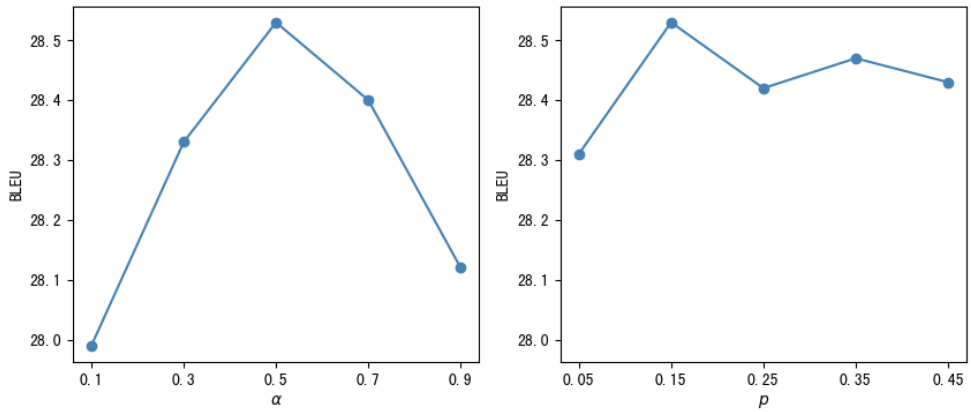

图 2. 在MuST-C英语-德语验证集上知识蒸馏权重α(左)、掩码概率p(右)对语音翻译性能的影响

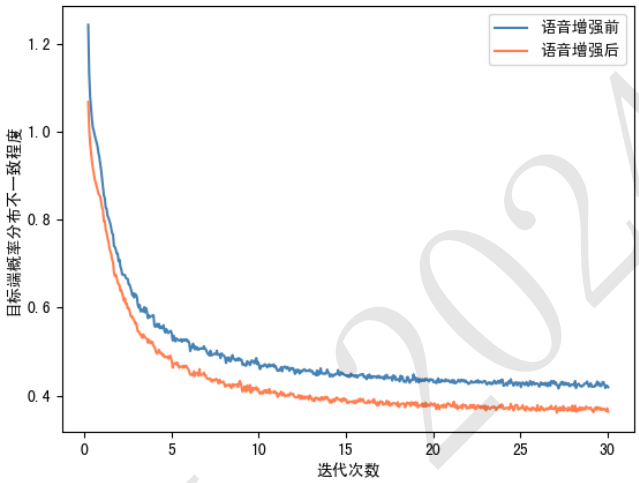

图 3. 在MuST-C英语-德语训练过程中目标端预测概率分布的不一致程度

码概率p对最终语音翻译性能的影响，本文同样在MuST-C英语-德语验证集上进行网格搜索，令$p \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$，实验结果如图2(右)所示。从图中可以看出，当$p = 0.15$时，MuST-C验证集上性能最佳。因此，本文其它实验均设置$\alpha = 0.5$，$p = 0.15$。

5.5 模态差异

本文利用公式(9)来衡量语音翻译任务和语音增强的文本翻译任务预测的目标端概率分布的不一致程度。为了探究语音增强前后目标端概率分布不一致程度的变化状况，本文在英语-德语方向的训练过程中绘制了语音增强前后的不一致程度变化曲线，如图3所示。其中，“语音增强前”表示模型在语音翻译任务、文本翻译任务和模态正则化损失上训练得到(即ST+TT)，“语音增强后”表示模型在语音翻译任务和本文提出的语音增强的文本翻译任务上训练得到(即ST+SATT)。从图中可以看出，在相同迭代次数下，进行语音增强后，语音翻译任务和文本翻译任务对应的目标端概率分布不一致性程度更低，进一步表明了语音增强的有效性。

为了判断本文提出的辅助任务是否缓解了语音和文本之间的模态差异，本文使用余弦相似度来衡量二者的模态差异。具体而言，针对语音翻译任务和语音增强的文本翻译任务，我们分别获取其经过翻译编码器后的输出$\mathbf{e_{mix}}$和$\mathbf{a}$，接着在时间维度进行平均池化来获取相应的语音表示和文本表示并计算二者的余弦相似度。本文在MuST-C英语-德语测试集tst-COMMON进行实验。实验结果如表6所示，表中各个模型代表的含义与5.3节相同。从表中#1和#2可以看出，对文本翻译任务进行语音增强后能够将语音表示和文本表示的余弦相似度提升2.45%。从

#	模型	余弦相似度（%）
1	ST+TT	86.20
2	ST+SATT	88.65
3	本文方法	**90.32**

表 6. 在MuST-C英语-德语测试集tst-COMMON上模态差异对比

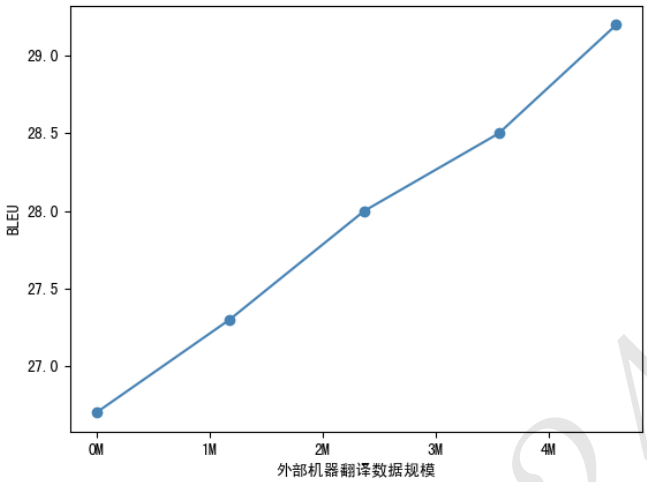

图 4. 在MuST-C英语-德语测试集tst-COMMON上外部机器翻译数据规模与性能对比

表中#2和#3可以看出，本文提出的全局感知条件掩码语言建模任务能够把余弦相似度再次提升1.67%。余弦相似度的提升验证了本文提出的辅助任务能够缓解语音和文本的模态间差异。

#	模型	BLEU
1	TT	**34.0**
2	ST	27.2
3	ST+SATT	33.5
4	本文方法	**34.0**

表 7. 在MuST-C英语-德语测试集tst-COMMON上的文本翻译任务性能对比

5.6 文本翻译性能

在多任务学习框架下，语音翻译任务性能的提升可能会导致文本翻译任务性能出现下降。为了探究本文提出的辅助任务是否会损害文本翻译任务的性能，本文在MuST-C英语-德语测试集tst-COMMON上进行实验。实验结果如表7所示，表中“TT”表示模型仅在文本翻译任务上训练，其它模型代表的含义与5.3节相同。从表中#1和#2可以看出，模型仅在语音翻译任务上训练后，文本翻译任务的性能出现大幅下降(-6.8)。从表中#1和#3可以看出，在应用本文提出的语音增强的文本翻译任务后，文本翻译任务的性能仅出现略微下降(-0.5)。最终，再加入全局感知条件掩码语言建模任务后，原本文本翻译任务的性能依然得到了保留。

5.7 外部机器翻译数据对语音翻译性能的影响

为了探究外部机器翻译数据对本文方法的影响，本文在MuST-C数据集的英语-德语方向进行了实验。将英语-德语方向的460M外部机器翻译数据分成五等份，并逐步增加外部机器翻译数据的使用量。针对每个增量量级分别在tst-COMMON测试集上进行评估。实验结果如图4所示。从图中可以看出，随着外部机器翻译数据使用量的增加，本文方法的BLEU值持续提高。

6 总结

本文在多任务学习框架下提出两个新颖辅助任务来改善语音翻译任务的训练。其中，语音增强的文本翻译任务通过使转录文本关注语音模态的信息，缓解了语音和文本之间的模态差异，从而提升了语音翻译任务的性能。全局感知的条件掩码语言建模任务使模型能够同时建模转录文本和译文，学习文本自身的全局上下文信息并利用这些信息指导翻译模型的训练。相比于其它最新的端到端语音翻译模型，本文的方法能够在MuST-C数据集上取得有竞争力的性能。相关的分析实验表明本文方法的有效性，它能够缓解语音和文本之间的模态差异，同时不损害文本翻译任务的性能。

参考文献

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NIPS*, volume 33, pages 12449–12460.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On using specaugment for end-to-end speech translation. In *Proceedings of IWSLT*.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of ACL*, pages 7893–7905, July.

Nanxin Chen, Shinji Watanabe, Jesús Villalba, Piotr Żelasko, and Najim Dehak. 2021. Non-autoregressive transformer for speech recognition. *IEEE Signal Processing Letters*, 28:121–125.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023. M3st: Mix at three levels for speech translation. In *Proceedings of ICASSP*, pages 1–5.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of NAACL*, pages 2012–2017.

Qianqian Dong, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu, and Bo Xu. 2019. Adapting translation models for transcript disfluency detection. In *Proceedings of AAAI*, pages 6351–6358.

Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. In *Proceedings of ACL*, pages 15864–15881.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of ACL*, pages 7050–7062.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of EMNLP-IJCNLP*, pages 6112–6121.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of ACL-IJCNLP*, pages 2214–2225.

Yuchen Han, Chen Xu, Tong Xiao, and Jingbo Zhu. 2023. Modality adaption or regularization? a case study on end-to-end speech translation. In *Proceedings of ACL*, pages 1340–1348.

Yosuke Higuchi, Shinji Watanabe, Nanxin Chen, Tetsuji Ogawa, and Tetsunori Kobayashi. 2020. Mask CTC: Non-Autoregressive End-to-End ASR with CTC and Mask Predict. In *Proceedings of INTERSPEECH*, pages 3655–3659.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Wuwei Huang, Dexin Wang, and Deyi Xiong. 2021. AdaST: Dynamically adapting encoder states in the decoder for end-to-end speech-to-text translation. In *Findings of ACL-IJCNLP*, pages 2539–2545.

Hirofumi Inaguma, Yosuke Higuchi, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2021. Orthros: non-autoregressive end-to-end speech translation with dual-decoder. In *Proceedings of ICASSP*, pages 7503–7507.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *Proceedings of ICASSP*, pages 7180–7184.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2021. Cascaded models with cyclic feedback for direct speech translation. In *Proceedings of ICASSP*, pages 7508–7512.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of ACL*, pages 245–254.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL*, pages 48–53.

Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: Word-aligned contrastive learning for speech translation. In *Proceedings of ACL*, pages 3891–3907.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D. McCarthy, and Deepak Gopinath. 2019. Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade. In *Proceedings of IWSLT*.

Juan Miguel Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. In *Proceedings of INTERSPEECH*, pages 1476–1480.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. SimulSpeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of ACL*, pages 7409–7421.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of ACL-IJCNLP*, pages 4252–4261.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020a. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proceedings of AAAI*, pages 9161–9168.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021. Large-scale self-and semi-supervised learning for speech translation. In *Proceedings of INTERSPEECH*, pages 2242–2246.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. In *Proceedings of IJCAI*, pages 6796–6804.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-End Speech Translation via Cross-Modal Progressive Training. In *Proceedings of INTERSPEECH*, pages 2267–2271.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of NAACL*, pages 5099–5113.

Wenbiao Yin, Zhicheng Liu, Chengqi Zhao, Tao Wang, Jian Tong, and Rong Ye. 2023. Improving speech translation by fusing speech and text. In *Findings of EMNLP*, pages 6262–6273.

Pei Zhang, Niyu Ge, Boxing Chen, and Kai Fan. 2019. Lattice transformer for speech translation. In *Proceedings of ACL*, pages 6475–6484.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020a. Adaptive feature selection for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2533–2544.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020b. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of ACL*, pages 1628–1639.

Hao Zhang, Nianwen Si, Yaqi Chen, Zhen Li, Tong Niu, Xukui Yang, and Dan Qu. 2022a. FCGCL: Fine- and coarse-granularity contrastive learning for speech translation. In *Findings of EMNLP*, pages 3048–3059.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022b. SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of EMNLP*, pages 1663–1676.

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang, and Yaqian Zhou. 2023a. DUB: Discrete unit back-translation for speech translation. In *Findings of ACL*, pages 7147–7164.

Linlin Zhang, Kai Fan, Boxing Chen, and Luo Si. 2023b. A simple concatenation can effectively improve speech translation. In *Proceedings of ACL*, pages 1793–1802.

Yuhao Zhang, Chen Xu, Bei Li, Hao Chen, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023c. Rethinking and improving multi-task learning for end-to-end speech translation. In *Proceedings of EMNLP*, pages 10753–10765.

Chulun Zhou, Fandong Meng, Jie Zhou, Min Zhang, Hongji Wang, and Jinsong Su. 2022. Confidence based bidirectional global context aware training framework for neural machine translation. In *Proceedings of ACL*, pages 2878–2889.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of ACL*, pages 7873–7887.