基于端到端的藏汉语音翻译

步寅硕[1,2,3],仁增多杰[1,2,3],格桑加措[1,2,3],拉毛吉[1,2,3],尼玛扎西[1,2,3]
(1. 西藏大学 信息科学技术学院,西藏 拉萨 850000;
2. 西藏自治区藏文信息技术人工智能重点实验室,西藏 拉萨 850000;
3. 藏文信息技术教育部工程研究中心,西藏 拉萨 850000)

摘 要:语音翻译在跨语言交流中具有重要意义,它能够消除语言障碍,实现不同语言间的即时沟通。传统的级联式 S2ST 系统存在错误复合和延迟较高的问题。相比之下,端到端模型通过简化处理流程,有效减少延迟并提升翻译准确性。目前,端到端语音翻译的研究主要集中在高资源语言对,而在低资源语言对,尤其是藏汉语音翻译领域缺少相关研究成果。针对该问题,该文提出一种端到端藏汉语音翻译方法。该方法首先引入基于声学特征扰动增强方法的语音数据增强技术,解决藏汉语音翻译数据资源匮乏的问题。其次,引入双边扰动技术对 Hubert 模型进行微调,通过风格归一化和信息增强阶段减少声学多模态对翻译的影响。再次,引入 S2UT(Speech-to-units)模型实现源语言语音到目标语言离散单元的转换,以解决 Mel-spectrogram 映射中存在的语言内容与声学特征混淆的问题。最后,在模型中加入目标语言的语音识别辅助任务,通过联合解码提高语音翻译性能。实验结果显示在藏语-汉语语音翻译任务中,BLEU 分数相比基线模型提升了 12.61,验证了该模型在低资源多模态语言对下的有效性。

关键词:端到端;语音翻译;低资源;藏语;Hubert
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2025)06-0166-09
doi:10.20165/j.cnki.ISSN1673-629X.2025.0042

Tibetan-Chinese Speech-to-speech Translation Based on End-to-end

BU Yin-shuo[1,2,3],Renzeng Duojie[1,2,3],Kalzang Gyatso[1,2,3],Lhamao Kyi[1,2,3],Nyima Trashi[1,2,3]
(1. School of Information Science and Technology,Tibet University,Lhasa 850000,China;
2. Key Laboratory of Tibetan Information Technology and Artificial Intelligence of Tibet,Lhasa 850000,China;
3. Engineering Research Center of Tibetan Information Technology,Ministry of Education,Lhasa 850000,China)

Abstract:Speech-to-speech translation is of great significance in cross-language communication. It can eliminate language barriers and achieve instant communication between different languages. The traditional cascade S2ST system has the problems of error compounding and high latency. In contrast,the end-to-end model effectively reduces latency and improves translation accuracy by simplifying the processing flow. At present,the research on end-to-end speech-to-speech translation mainly focuses on high-resource language pairs,while there is a lack of relevant research results in low-resource language pairs,especially in the field of Tibetan-Chinese speech-to-speech translation. Therefore,we propose an end-to-end Tibetan-Chinese speech-to-speech translation method. Firstly,the speech data enhancement technology based on the acoustic feature perturbation enhancement method is introduced to solve the problem of scarce data resources for Tibetan-Chinese speech-to-speech translation. Secondly,the bilateral perturbation technology is introduced to fine-tune the Hubert model,and the influence of acoustic multimodality on translation is reduced through style normalization and information enhancement stages. Thirdly,the Speech-to-units (S2UT) model is introduced to realize the conversion of source language speech to target language discrete units to solve the problem of confusion between language content and acoustic features in Mel-spectrogram mapping. Finally,the speech recognition auxiliary task of the target language is added to the model to improve the speech-to-speech translation performance through joint decoding. The experimental results show that in the Tibetan-Chinese speech-to-speech translation task,the BLEU score is improved by 12.61 compared with the baseline model. The results demonstrate the effectiveness of the proposed model in low-resource multimodal language pairs.

Key words:end-to-end;speech-to-speech translation;low resources;Tibetan;Hubert

收稿日期:2024-11-12 **修回日期**:2025-03-13
基金项目:新一代人工智能国家科技重大专项(2022ZD0116100);西藏大学研究生“高水平人才培养计划”项目(2022-GSP-S101)
作者简介:步寅硕(2000-),女,硕士研究生,研究方向为自然语言处理;通讯作者:尼玛扎西(1964-),男(藏族),教授,研究方向为自然语言处理、认知智能。

0 引　言

在多元文化的交流中,语音到语音翻译(Speech-to-speech Translation,S2ST)技术显得尤为重要,其宗旨在于实现不同语言间的语音直接转换,促进了跨语言界的沟通。传统的语音到语音翻译(S2ST)系统通常由三个组件组成:自动语音识别(Automatic Speech Recognition,ASR)、文本到文本的机器翻译(Machine Translation,MT)和文本到语音合成(Text To Speech,TTS)。通过级联来实现语音到语音的翻译[1-3]。然而,这种分步处理方式往往因各环节错误的累积而影响最终翻译质量。近年来,端到端 S2ST 模型的出现为解决这一问题提供了新思路。相较于传统方法,端到端模型在转换过程中能够更好地保留语音的原始特质,如说话人的声音和语调,且对于无书面形式的语言也能进行有效处理。此外,该模型在计算效率和推断速度上具有明显优势,减少了翻译过程中的错误积累,并能更加灵活地处理特定内容,例如不需翻译的专有名词。综上所述,端到端的 S2ST 模型不仅技术先进,更在实际应用中展现出其对于破除语言壁垒、促进全球互联互通的巨大潜力。

到目前为止,端到端语音到语音翻译已经取得了巨大的进步。Translatotron[4]是第一个端到端 S2ST 模型,它通过直接将源语言的语音谱图映射到目标语言的语音谱图上,标志着端到端语音翻译成为可能。Takatomo Kano 等人[5]的工作强调了在处理结构相似或差异显著的语言对时,基于 Transformer 的架构相较于传统的 RNN 架构,展现出了更优的适应性和效率。Dong Qianqian 等人[6]通过引入 S2ST Transformer 模型,将 LSTM 替换为 Transformer,为该领域的技术进步提供了新的动力。Lee 等人[7]提出了端到端 S2ST 系统,利用了自监督学习(Self-Supervised Learning,SSL),无需依赖文本数据即可实现翻译。此外,Wei 等人[8]构建了 Speech2S,利用未配对语音和双语文本数据进行预训练,在数据稀缺条件下有效提升了翻译性能。然而,大多数 SSL 模型都是通过重构[9]或预测看不见的语音信号[10]来训练的,这使得其中不可避免地会包括与语言内容无关的因素(即声学条件)。因此,语音到单元的翻译中,训练目标的不确定性将导致无法产生良好的结果。无文本 S2ST 系统[11]进一步演示了通过微调 SSL 模型来解开依赖于说话人的信息,从而获得说话人不变的表示。然而,该系统仅约束说话人身份,而其余方面(即内容、节奏、音高和能量)仍然集中在一起。Huang Rongjie 等人[12]提出模型 TranSpeech,证明了使用扰乱信息流以微调声学模型的方法能够促进下游任务的性能。

此外,由于 S2ST 缺乏训练数据,之前关于该方向的研究的大部分工作都是在 S2UT(Speech-to-units)语料库上应用 TTS 生成用于模型训练的合成目标语音的数据集上进行实验的[13-14]。此前关于端到端语音翻译的研究主要集中在高资源和相似语言对方面,如英语与西班牙语之间的语音翻译,取得了一定的研究进展[11],但小语种的翻译工作目前并未有相关工作,S2ST 在低资源设置下对远距离语言对的可行性仍然未知。

在藏汉语音翻译领域,目前研究主要采用级联方法,大部分工作集中于“语音-文本”的翻译。例如,结合语音识别与机器翻译技术实现藏汉翻译[15],以及通过多阶段处理提升语音转换的音质和自然度[16]。然而,针对端到端的藏汉语音翻译研究仍处于空白阶段。这表明,如何实现低资源、远距离语言对的端到端语音翻译,成为当前亟待解决的重要问题。

到目前为止,端到端语言翻译已经取得了巨大进步。但目前 S2ST 系统的发展面临两大阻碍:一是数据稀缺,缺乏并行语音数据对,导致现有模型无法充分利用稀缺数据有效收敛,模型结果较差;二是由于声学多模态(如说话者身份、节奏、音高和能量),很难实现高翻译精度。因此,如何解决语音翻译中的数据稀缺问题以及多模态问题,有效提高翻译精度,成为当前亟待解决的问题。

基于上述问题,研究主要从以下三个方面进行。首先,针对藏汉语言对数据稀缺的情况,探索如何有效利用现有数据,通过声学特征扰动增强技术扩展数据集,解决数据资源匮乏的问题。其次,针对声学多模态挑战,研究如何处理说话者身份、音高、节奏等多模态信息,以提升端到端语音翻译系统在不同语音特征下的稳定性和准确性。最后,探究端到端语音翻译模型在小语种翻译工作中的可行性,如藏汉这种低资源的远距离语言对之间的语音翻译,通过在模型中加入目标语言的语音识别辅助任务,利用联合解码以提高语音翻译性能。

1 端到端语音翻译模型

该文利用基于离散单元的 S2ST 方法[7],通过自监督语音编码器技术,实现目标语音向整数序列的转化,并将源语音映射为目标离散单元,进而构建了一套藏语-汉语的翻译体系。该模型架构详见图 1。

为了支持低资源多模态语言对藏语→汉语的翻译,首先基于信息增强技术扩展原有数据集,解决语音数据资源匮乏的问题。其次,针对声学多模态挑战,扩展了基于 Hubert 的离散单元提取技术[17],优化了预训练声学模型[12],旨在提炼出更加准确、与声学属性(如说话者特征、节奏、音高及能量)无关的声学表示。此

外,该文选用了 S2UT 架构[7],实现了源语言音频直接转换为目标语言离散单元,并通过引入辅助任务,对于具有书面语形式的汉语这一目标语言,进一步采用了基于离散单元解码器中间表示的 CTC 目标文本解码策略,实现了语音与文本的联合训练与生成,有效提升了模型的转换效率与准确性。模型最后进行了声码器的独立训练,从而成功实现了从离散单元到波形的转换[18],验证了该方法在汉语领域的应用潜力。下面将对上述技术进行详细介绍。

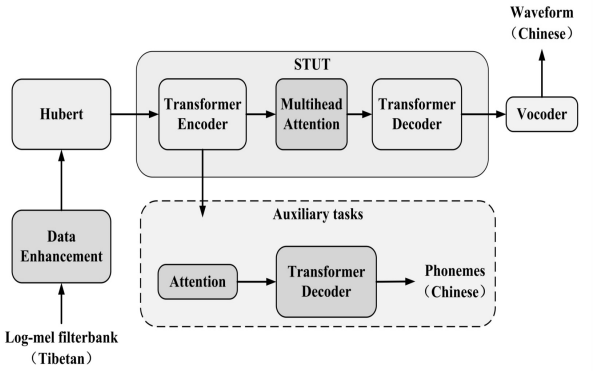

图 1 模型架构

1.1 声学特征扰动增强技术

针对当前缺乏真实的藏汉语音翻译数据集且现有的弱监督数据集数据量不足的问题,该文提出一种数据增强的方法,以扩大现有数据集的规模,提高数据集的多样性,从而更有力地支持藏汉语音翻译研究的深入开展。原始语音数据集为 $S=\{x_1,x_2,\cdots,x_N\}$,其中每个 x_i 表示语音样本。在原始的藏汉语音翻译数据集基础上,引入了声学特征扰动增强的方法,对原始语音样本中的说话人身份、节奏、音高和能量进行变化,从而将原始数据集扩充至五倍。

在这一过程中,使用了四种不同的函数,分别作用于不同的声学特征,从而在保持语言内容一致的情况下生成声学特征发生变化的样本。这些函数包括:共振峰偏移(formant shifting,FS)、音高随机化(pitch randomization,PR)、随机频率塑形(random frequency shaping using a parametric equalizer,PEQ)以及随机重采样(random resampling,RR)。

(1)formant shifting(FS)函数:用于调整语音的共振峰,通过从均匀分布 Unif(1,1.4)中采样共振峰偏移比率实现。采样后随机决定是否取该比率的倒数。这种微调语音频谱特征的方法可以在不改变语言内容的前提下使音质发生轻微变化。

对每个语音样本 x_i 的共振峰频率 $F(x_i)$ 进行扰动,生成新的样本。

$$F'(x_i)=F(x_i)\bullet R_f,R_f\sim \mathrm{Unif}(1,1.4)\cup\{1/R_f\} \tag{1}$$

其中,R_f 为均匀分布采样的偏移比率,随机选择是否取其倒数。

(2)pitch randomization(PR)函数:通过对音高比率进行调整来引入音高的随机变化。音高偏移比率和音高范围比率分别从均匀分布 Unif(1,2)和 Unif(1,1.5)中采样,并再次随机选择是否取采样比率的倒数。

对每个语音样本 x_i 的音高比率 $P(x_i)$ 和音高范围 $R(x_i)$ 进行扰动,生成:

$$P'(x_i)=P(x_i)\bullet R_p,R_p\sim \mathrm{Unif}(1,2)\cup\{1/R_p\} \tag{2}$$

$$R'(x_i)=R(x_i)\bullet R_r,R_r\sim \mathrm{Unif}(1,1.5)\cup\{1/R_r\} \tag{3}$$

其中,R_p 和 R_r 分别表示音高调整的扰动因子和音高范围的扰动因子,随机选择是否取其倒数。

(3)random frequency shaping using a parametric equalizer(PEQ)函数:由低搁架滤波器(low-shelving,HLS)、峰值滤波器(peaking,HPeak)和高搁架滤波器(high-shelving,HHS)组成。通过使用一个低搁架滤波器、一个高搁架滤波器和八个峰值滤波器,对语音的频率响应进行调节,从而增加语音频谱的多样性。

频率塑形通过调整语音样本 x_i 的频率响应 $G(x_i,f)$ 实现,计算公式如下:

$$G'(x_i,f)=G(x_i,f)\bullet H(f) \tag{4}$$

其中,$H(f)$ 为滤波器增益函数,定义为:

$$H(f)=H_{\mathrm{HLS}}(f)\bullet\prod_{j=1}^{8}H_{\mathrm{HPeak},j}(f)\bullet H_{\mathrm{HHS}}(f) \tag{5}$$

其中,$H_{\mathrm{HLS}}(f)$ 表示低搁架滤波器,用于调节低频部分的增益,对低频能量进行增强或衰减。$H_{\mathrm{HPeak},j}(f)$ 表示峰值滤波器,包含八个独立的峰值滤波器,分别用于对中频段的特定频率进行精细调整,增加频谱的多样性。$H_{\mathrm{HHS}}(f)$ 表示高搁架滤波器,用于调节高频部分的增益,对高频能量进行增强或衰减。

(4)random resampling(RR)函数:用于对节奏信息进行随机重采样。将输入信号分为随机长度的段落,每段长度在 19 帧到 32 帧之间均匀采样。每段通过随机选择的重采样因子(范围为 0.5 到 1.5)进行线性插值重采样,实现对每个段落的时间拉伸或压缩,以生成具有时间变化的语音样本。具体步骤如下:

首先将语音样本 $x_i\in S$ 分割为 m 段,表示为:$\{x_{i,1},x_{i,2},\cdots,x_{i,m}\}$,每段表示语音的连续时间片段。

然后对每个片段 $x_{i,j}$ 进行随机重采样,调整其时间长度,公式如下:

$$x'_{i,j}=\mathrm{Resample}(x_{i,j},r_j),r_j\sim \mathrm{Unif}(0.5,1.5) \tag{6}$$

其中,r_j 是从均匀分布 Unif(0.5,1.5)中随机采样的重采样因子,用于拉伸或压缩该段的时间维度。

将所有重采样后的段拼接,生成新的语音样本 $x_i^{'}$。

$$x_i^{'} = \bigcup_{j=1}^{m} x_{i,j}^{'} \tag{7}$$

扰动后的语音在声学特征(即节奏、音高和能量)方面展现出显著的变化性,数据增强技术通过对节奏、音高及频率分布的扰动,生成了与原始样本不同的特征分布样本,同时确保了语言信息的完整性,保留了语义特征。

这些增强策略在拓展多样性的同时,对分布规律产生了适度的改变。

(a)节奏分布的调整:增强后的样本覆盖了更多快语速和慢语速的语音特征,减少了原始数据中语速分布的偏差。

(b)音高分布的扩展:通过随机增加或减少音高范围,增强技术生成了更多高音高和低音高的样本,使音高分布更加均匀。

(c)能量特性的多样化:能量扰动增加了样本中不同能量级别的分布比例,尤其是在低能量和高能量区域,拓展了覆盖范围。

因此,数据增强技术在扩展数据集的同时,适度改变了分布规律,使数据集从原本可能存在的集中分布变得更加均衡和多样化。此外,针对语音的节奏特性进行的变化,有效缓解了数据集中某些特征稀缺的问题,从而进一步扩大了数据集的覆盖范围,为模型的鲁棒性和泛化能力提供了强有力的支持。

1.2　基于双边扰动的多模态语音风格归一化

自监督预训练模型得出的语音表示融合了丰富的语言及声学信息[11]。然而,这种语音表示受到声学信息变化(即声学多模态)的影响,导致相同语音内容的样本推导出不同的表示(参见图 2),这种语音至单元转换的训练目标的不确定性将导致模型训练难以取得满意成效。针对这一复杂性问题,该文采用基于单元的微调策略,对预训练的声学模型进行了调整,以此对 Hubert 模型进行改进和优化。

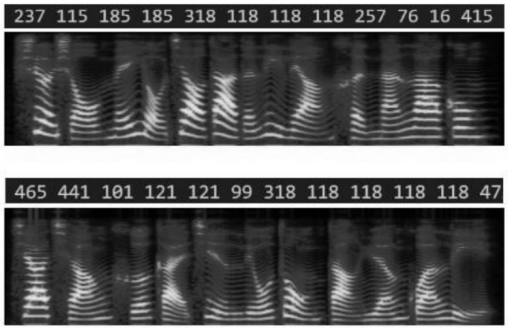

图 2　不同声学条件的离散单元

为了深入探究语音变化[19],该文将其细分为两大要素:语言内容与声学条件。将声学条件进一步划分为讲话者身份、节奏、音高及能量等关键因素。其中语言内容指的是语音信号中传达的具体语义信息,包括词汇、句法结构和表达的意义,构成了语音交流的基础,使听者能够理解讲话者的意图和信息。讲话者身份指的是语音中包含的能够辨别出讲话者的独特特征,如音色、语速和口音等。它使得同样的语言内容在不同人之间呈现出个性化的表达。节奏描述了讲话者在发音过程中音节的分布模式,即语速的快慢和停顿的安排。节奏通过音节的时长和重音分布形成了独特的发声节拍。音高指的是声音的频率高低,是语音信号中实现音节差异的方式。通过音高的不同变化,可以使音节在句子中形成升降的音调。能量是语音的响度表现,反映了语音信号的强度。在语音中,不同的能量值通过重音和声调的变化体现出来,影响着语音信号的总体音量和听觉效果。

为了实现对 Hubert 模型的微调,该文进行了双边扰动处理,具体而言分为两个部分:首先是针对原始数据集生成“伪文本”,其次是在原始数据集的基础上生成扰动语音。

1.2.1　生成“伪文本”

为了在连接时序分类(CTC)模型的目标数据中去除那些与语音信号的声学特征相关的信息,如讲话者的身份、节奏、音高和能量等,通过预训练 SSL 模型创建与声学变化无关的“伪文本”,该文本只包含纯粹的语言内容,而不带有任何声学特征,从而为模型提供一个更加稳定和一致的训练目标。该阶段分为三步进行:

(1)对原始数据集 $S = \{x_1, x_2, \cdots, x_N\}$ 的平均基频 $\bar{F}_0$ 和能量 $\bar{E}$ 值进行计算。

$$\begin{aligned} \bar{F}_0 &= \frac{1}{N}\sum_{i=1}^{N} F_0(x_i) \\ \bar{E} &= \frac{1}{N}\sum_{i=1}^{N} E(x_i) \end{aligned} \tag{8}$$

其中,$F_0(x_i)$ 和 $E(x_i)$ 分别表示语音样本 x_i 的基频和能量。

(2)对于 S 中的每个样本,进行音高移位到 $\bar{F}_0$,并将其能量标准化为 $\bar{E}$,通过以下公式实现。

$$\begin{aligned} F_0^{'}(x_i) &= F_0(x_i) - (F_0(x_i) - \bar{F}_0) \\ E^{'}(x_i) &= E(x_i) - (E(x_i) - \bar{E}) \end{aligned} \tag{9}$$

调整后,可以得到一个具有平均声学条件的新数据集 $S_1 = \{x_1^{'}, x_2^{'}, \cdots, x_N^{'}\}$,此时,原始数据集中的关于声学变化的特定信息已经被消除,每个样本在基频和能量上具有一致的声学特征。

(3)自监督学习(SSL)模型对 $S_1 = \{x_1^{'}, x_2^{'}, \cdots, x_N^{'}\}$

进行编码,并创建用于 CTC 微调的标准化目标“伪文本” T ,公式如下:

$$T = \mathrm{SSL}(S_1) \tag{10}$$

其中, T 是一种去除了声学特征变化的文本表示,包含纯粹的语音内容信息,作为 CTC 模型的标准化目标。

1.2.2 生成扰动语音

通过信息瓶颈技术对输入语音数据进行处理,以提取任务相关的关键信息,并有效过滤掉与任务无关的冗余信息。这一方法能够在保留语音信号中最有价值内容的同时,减少不必要的干扰。在这里通过特定函数对语音样本进行处理,具体函数如下:

(1)谱域转换(spectral transformation, ST)函数:该函数通过在傅里叶变换域内对频域幅度引入少量随机噪声,实现音色的变化。具体操作为在频域中对信号的幅度进行随机扰动,并在逆变换过程中恢复音色的变化效果。

具体公式为:

$$S'(f) = S(f) \bullet G(f) \tag{11}$$

其中, $S(f)$ 为原始语音的频谱, $G(f)$ 为频率增益函数,用于模拟不同语音音色的变化。通过将 $G(f)$ 的范围控制在 [- 3 dB, + 3 dB] 之间,使处理后的语音样本展现出不同的说话人特征。

(2)非线性动态音高调整(nonlinear dynamic pitch adjustment, NDPA)函数:NDPA 函数通过分段非线性调整音高,以生成多样化的音高变化模式。该方法将语音信号划分为等长片段,并依据随机生成的音高变换公式在各片段上施加不同的非线性变化,从而增强音高的多样性,模拟出自然语音中常见的音高波动,公式为:

$$P'(t) = P(t) + A \bullet \sin(2\pi f t) \tag{12}$$

其中, $P(t)$ 为原始音高, A 为扰动幅度, f 为正弦波的频率, t 为时间。

(3)频率选择扰动(frequency selection perturbation, FSP)函数:此函数通过分段选择性滤波,将语音信号的特定频段在不同时间段内进行衰减或增强。FSP 函数随机选择低频和中高频区间,并在各区间内引入少量白噪声或提升幅度,以形成多种频率变化效果,丰富了语音样本的频谱特征。

通过分段选择频段对语音频率特性进行增强或削弱,调整后的频谱为:

$$S'(f) = S(f) \bullet M(f) \tag{13}$$

其中, $M(f)$ 为频率掩蔽函数,用于随机抑制或增强不同频段的幅度,增强语音的频谱变化特性。

(4)时序变换(temporal transformation, TT)函数:TT 函数基于自适应时间拉伸和压缩技术对语音的节奏特征进行扰动。具体方法为将信号划分为多个段落,对各段落的采样率在[0.7,1.3]范围内随机变化。为保持节奏变换的自然连贯性,TT 函数在各段落间采用平滑过渡的时序调整。公式如下:

$$t' = t \bullet r, r \sim \mathrm{Unif}(0.7, 1.3) \tag{14}$$

其中, r 为随机采样的时间伸缩因子,能够拉伸或压缩语音节奏,使语音在时间维度上呈现变化。

(5)能量动态扰动(energy dynamic perturbation, EDP)函数:EDP 函数通过周期性调整语音信号的能量分布来引入不同的能量变化层次,模拟自然语音中的音量波动。该方法在时间轴上随机选择多个节点,并以正态分布的扰动幅度调整每个节点的音量,生成自然的能量起伏特征。

公式如下:

$$E'(t) = E(t) \bullet \alpha \tag{15}$$

其中, $\alpha \sim \mathrm{Unif}(0.8, 1.2)$ 为随机选取的能量调节因子,用于增加语音的能量动态变化。

在每次处理过程中,以上方法被随机组合使用,每条原始语音样本 x_i 均经过上述一种或多种方法的处理,从而生成一条对应的扰动语音样本 x_i' 。通过以上函数组合,对语音的频域、时域、能量和音高特征实施多维度扰动,使得语音的声学特征(说话人身份、节奏、音高和能量)发生显著变化,但同时保持语音的语言内容不变,从而生成一组新的语音样本 S_2。

1.2.3 双边扰动

针对多说话人这种更加真实的场景下的声学变化,为了缓解 S2ST 系统中的声学多模态问题并提高翻译精度,该文利用了连接主义时间分类(CTC)微调技术[20],结合预训练的语音编码器,通过引入扰动输入语音和目标语音的规范化处理,有效地解决了声学特征中的多模态困扰,从而根据语言内容生成确定性的语音表示。

双边扰动可以看作是使用扰动语音 S_2 作为输入,“伪文本”作为目标来训练 ASR 模型,通过对 Hubert 进行微调使得具有 CTC 解码的模型学习涉及语言内容的“平均”信息并生成确定性表示,显著减少了声学多模态。

在多说话人场景下,选择某一说话人作为参考,其他说话人与参考说话人的内容是相同的,用 S 表示原始语音数据集,用 S_1 和 S_2 分别表示风格归一化和信息增强后的语音样本。因此源语言是一个语音样本序列 $X = \{x_1, x_2, \cdots, x_{N'}\}$,其中 N' 为源语音的帧数。SSL 模型由多层卷积特征编码器组成,该编码器以原始音频 S 为输入,输出离散的潜在语音表示。最后,将目标语言中的音频表示为离散单元 $Y = \{y_1, y_2, \cdots, y_N\}$,其中 N' 为单元数。加入双边扰动后对 SSL 模型进行微调的流程如图 3 所示。

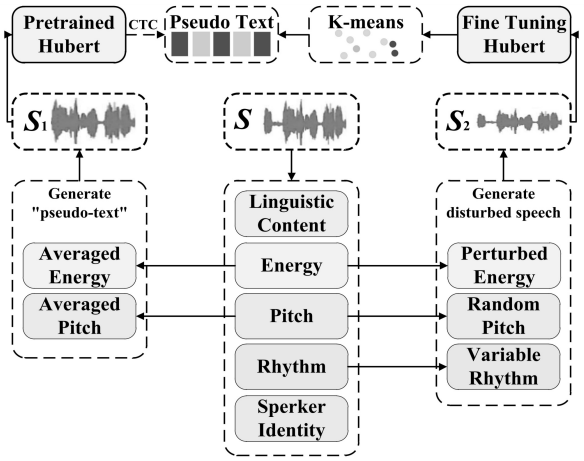

图 3　双边扰动流程

1.3　基于 S2UT 模型的源语音到目标离散单元转换

为了实现端到端语音翻译,目前较有效的方法是实现语音到离散单元的映射。为此,引入了 S2UT (Speech-to-units)模型,以实现源语言语音到目标语言离散单元的转换。具体而言,通过改编自机器翻译中的 Transformer 模型[21],构建了 S2UT 模型,该模型由语音编码器和离散单元解码器组成。在编码器中,为了实现对语音进行下采样[22],从而在保留重要信息的同时减少输入的时间步数或特征数,在编码器前添加了堆叠的 1 维卷积层。假设输入的语音信号为 x,经过 L 层 1 维卷积后,时间步数为原来的 $1/2^L$。公式如下:

$$x' = \mathrm{Conv1D}_L(\mathrm{Conv1D}_{L-1}(\ldots \mathrm{Conv1D}_1(x))) \quad (16)$$

其中每层 1 维卷积的步幅均为 2。这种下采样操作有效减少了模型的计算复杂度,为后续的编码任务提供了更加精简的输入表示。

在选择预测离散单元序列的策略时,选择“reduced”方法[11],该方法将一系列相同单元的连续序列合并为一个单一单元,从而得到一系列唯一的离散单元。例如,假设目标语言的离散单元序列为 $u = \{u_1, u_2, u_2, u_3, u_3, u_3, u_4\}$,通过“reduced”方法,将其合并为 $u' = \{u_1, u_2, u_3, u_4\}$,其中,$u'$ 表示合并后的唯一离散单元序列。这一策略通过减少目标序列的长度,降低了解码复杂度,从而加速了推理过程。

由于目标语言是汉语,属于存在书面系统的语言,因此在 S2UT 基线模型的基础上,该文在 S2UT 模型中加入了辅助任务。提取目标语言的音素,并将其作为辅助任务的目标输出。该辅助任务只在训练期间使用,并不在推理过程中使用。在训练期间,使用真实目标离散单元序列 $Y = \{y_1, y_2, \cdots, y_T\}$ 进行教师强制,并使用来自解码器的教师强制中间表示计算 CTC 损失。CTC 损失定义为:

$$L_{\mathrm{CTC}} = -\log \sum_{a \in A(y)} P(a \mid x) \quad (17)$$

其中,$A(y)$ 表示所有可能的对齐路径,$P(a \mid x)$ 表示对齐路径 a 的概率。

在推理过程中,可以在每个解码步骤同时对文本进行离散单元解码和 CTC 解码。通过加入辅助任务,能够提高 S2UT 模型的性能。

此外,为了缓解语音和文本输出之间的长度不匹配问题,在模型中添加了 CTC 解码。然而,由于它只允许单调对齐,因此利用 CTC 解码器所依赖的 Transformer 层来处理从源语言到目标语言的重新排序。在推理期间同时进行离散单元解码和 CTC 解码。

最后,该文进行了声码器的独立训练,选择基于离散单元的 HiFi-GAN 声码器[23],并通过 Fastspeech 2[24] 中的持续时间预测模块进行了增强。在训练声码器时,使用通过双边扰动技术微调后的中文预训练 Hubert 模型所推理出的离散单元作为输入,训练过程中不使用额外的音高信息作为输入。该声码器的训练与 S2UT 模型分开进行,结合了 HiFi-GAN 中的生成器-判别器损失(generator-discriminator loss)和每个单元在对数域中预测持续时间的均方误差。

2　实　验

2.1　实验设置及评价指标

2.1.1　实验设置

对于双边扰动,使用 CTC 损失微调在 WenetSpeech[25] 上预训练的 Chinese Hubert Base 模型,直到进行 25 000 次更新。使用 k-means 算法将微调良好的 Hubert 模型第 9 层所给出的表示聚类成 500 个单元的词汇表。

在训练 S2UT 模型时,CTC 损失的权重设为 1.6。使用 Adam 模型训练 400k 步,$\varepsilon = 10^{-8}$,$\beta_1 = 0.9$,$\beta_2 = 0.98$,标签平滑 0.2,并应用平方根倒数学习率衰减计划,预热步数为 10k。所有其他超参数,如 dropout 和学习率,都在开发集上进行调优。

2.1.2　评价指标

在收集的藏汉双语语音数据(拉萨话-普通话)上进行训练和测试,使用训练集对选择的端到端模型进行训练,优化模型参数以最大化语音翻译性能。使用验证集对训练好的模型进行评估,对语音输出应用 ASR 并计算 BLEU 分数,以 BLEU 分数作为评估语音翻译文本质量的衡量标准,从而衡量语音质量。BLEU 分数的计算公式如下:

$$x = y_{\mathrm{BP}} \cdot \left(\sum_{n=1}^{N} w_n \cdot \log\ p_n\right) \quad (18)$$

其中:x 表示 BLEU 分数;y_{BP} 是长度惩罚因子,用于平衡生成翻译与参考翻译的长度差异;p_n 是第 n - gram 的精确率;w_n 是权重,通常均分为 $1/N$。长度惩罚因

子定义为：

$$y_{BP}=\begin{cases}1, & 若\ c>r\\ e^{(1-r/c)}, & 若\ c\leqslant r\end{cases} \tag{19}$$

其中，c 和 r 分别为生成翻译和参考翻译的总长度。表中 BLEU 分数以百分比表示，便于对不同模型性能的直观对比。

2.2 数据集

相比于传统级联系统，端到端语音翻译（S2ST）研究面临严重的数据稀缺问题，缺乏大规模的并行语音数据对。目前，端到端语音翻译的研究主要集中在高资源和语言结构相似的语言对，如西班牙语、英语等。这些语言对已具备公开的并行语音数据，为研究提供了充足的数据支持。然而，该文选择了藏语-汉语的端到端语音翻译作为研究对象，该语言对尚无公开的并行语音翻译数据集，现有的藏语数据仅限于语音识别领域。这一数据缺口增加了直接构建语音到语音翻译模型的挑战。为解决此问题，该文提出利用伪标签数据的方式构建藏汉语音翻译原始数据集，以缓解数据不足的现状，从而为端到端藏汉语音翻译模型的开发提供更为充分的数据支持。

使用的数据集来源有两部分：一部分是开源的数据集 TCST[26]，该数据集包含 7 个小时的拉萨话语音以及对应的汉语文本；另一部分是构建的数据集 Tibet-TXL，该数据包含了 25 个小时的藏语语音识别数据，其中仅包括藏语语音及对应的藏语文本。

针对 TCST 数据集，由于其中已经包含了藏语语音及其对应的汉语文本，处理相对直接。利用语音合成技术（TTS）将对应的汉语文本转换为与藏语语音匹配的汉语语音，以创建更丰富的多语言语音数据供模型训练使用。

针对 Tibet-TXL 数据集，由于该数据集中只包含藏语语音以及对应藏语文本，为了得到符合需求的藏语-汉语语言对，首先使用机器翻译技术（MT），将藏语文本翻译为中文文本，确保生成与目标语言一致的文本内容。随后，利用语音合成技术（TTS）将这些中文文本合成为中文语音，以便用于后续的模型训练和评估。为了提高效率，应用伪标签技术来创建弱监督数据，创建弱监督数据的流程详见图 4。

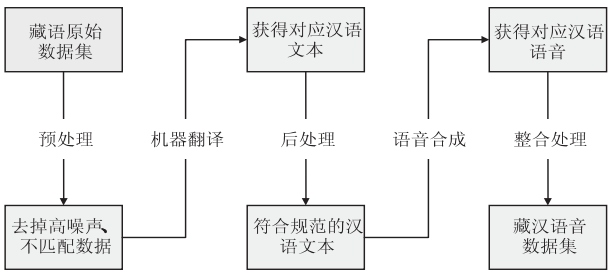

图 4 创建弱监督数据流程

数据集样本特点及分布情况：数据集中的语音样本涵盖了多种节奏、音高和能量特性，表现出如下规律：

（1）节奏：样本的语速分布可能集中于中速语音，快语速或慢语速的样本较少，呈现一定的不均衡性。

（2）音高：音高分布受说话人（如性别和年龄）影响明显，低音高（男性说话人）样本较多，高音高（女性或儿童说话人）样本占比较低。

（3）能量：语音样本的能量分布可能随语音内容和录音条件波动，存在高能量语音（如大声讲话）样本较少的情况。

通过伪标签技术成功构建了一个包含 30 小时藏语-汉语语音翻译的初始数据集。在此基础上，进一步利用上文所介绍的声学特征扰动增强技术对初始数据集进行了扩展和优化。具体来说，分别对藏语和汉语的语音样本施加了节奏、音高、能量等多维声学特征的扰动，生成了一系列具有声学多样性的语音数据。通过声学特征扰动增强技术，将原始数据集进行数据增强，从而得到新的数据集 Tibet-AT，该数据包含 148 小时的藏语语音数据（见表 1）。

表 1 数据集 h

	TCST	Tibet-TXL	原始数据集（30h）			Tibet-AT（148h）		
			Train	Valid	Test	Train	Valid	Test
藏语	7	25	29	1.5	1.5	145	1.5	1.5
中文	5.5	23	26.7	1	0.8	134	1	0.8

2.3 实验结果及分析

从表 2 中给出的实验结果数据可见，不同模型在 Tibet-AT 数据集上的 BLEU 评分差异较大。其中 Translatotron 模型通过语音频谱图的映射实现语音翻译，与语音到离散单元的映射相比，翻译性能更低。端到端 S2ST 系统相较于 Translatotron 有所改进，BLEU 评分达到了 3.11。TranSpeech 模型使用了扰乱信息流以微调声学模型，通过该方法，最终实现了 3.36 的 BLEU分值。作为对比实验的Speech2S模型，利用未

表 2 不同模型的性能对比

模型	BLEU			
	原始数据集（30h）		Tibet-AT（148h）	
	valid	test	valid	test
Translatotron[4]	0.43	0.35	1.26	1.19
端到端 S2ST 系统[7]	2.90	2.65	3.23	3.11
TranSpeech	\	\	3.53	3.36
Speech2S[8]	3.12	2.86	3.42	3.30
Ours	\	\	**16.31**	**15.72**

注：BLEU 分数以百分比形式（%）表示，即原始 BLEU 值 × 100%。

配对语音和双语文本数据进行预训练,在数据稀缺条件下表现出较好的性能,其 BLEU 评分为 2.86。该文提出的模型使用了更适合藏语-汉语语言对的 Hubert 模型,通过双边扰动来实现对 Hubert 模型的微调,并通过引入辅助任务来进一步提高模型性能,最终模型的 BLEU 评分达到了 15.72,证明了该模型在低资源语言对中的良好性能。

为了充分验证文中方法的有效性,分别对不同设置的模型进行了实验对比,评估其在原始数据集(30 小时)与数据增强后(148 小时)两种场景下的表现。通过逐步加入或移除关键模块(如 BIP 技术和辅助任务)来评估它们对模型性能的影响,通过逐步改变模型配置,验证不同模块对模型表现的贡献(参见表 3)。

表 3　消融实验

	BIP	辅助任务	BLEU			
			原始数据集(30h)		Tibet-AT(148h)	
			valid	test	valid	test
100 Hubert	×	×	2.90	2.65	3.23	3.11
100 Hubert	√	×	\	\	3.53	3.36
100 Huber	×	√	3.09	2.90	4.08	3.56
100 Hubert	√	√	\	\	8.04	6.85
500 Hubert	×	×	4.52	4.24	10.54	9.26
500 Hubert	√	×	\	\	12.30	11.05
500 Hubert	×	√	5.80	5.26	13.23	12.86
500 Hubert	√	√	\	\	**16.31**	**15.72**

注:BLEU 分数以百分比形式(%)表示,即原始 BLEU 值 × 100%。

实验首先在 100 Hubert 和 500 Hubert 模型上进行测试,未使用 BIP 技术和辅助任务。在原始 30 小时数据集上,基线模型的 BLEU 评分较低,分别为 2.65 以及 4.24。这一部分提供了模型的基础表现,用于后续比较。在基线模型的基础上,引入 BIP 技术。对于 100 Hubert 模型,引入 BIP 后,在数据增强后的 148 小时数据集上,BLEU 评分提升至 3.36。在 500 Hubert 模型中,BLEU 评分为 11.05,说明 BIP 技术能够有效提升模型的翻译性能。

在未使用 BIP 技术的情况下,仅加入辅助任务,可以观察到性能也有一定提升。对于 100 Hubert 模型,BLEU 评分进一步提升至 3.56。在 500 Hubert 模型中,BLEU 评分为 12.86,这表明辅助任务对模型的训练有积极作用,特别是在数据增强后的场景中。

在消融实验的最后阶段,模型同时引入了 BIP 技术和辅助任务,这是完整模型的表现。在这种配置下,模型性能达到了最佳。对于 500 Hubert 模型,数据增强后 BLEU 评分显著提升,达到 15.72。

通过上述消融实验,可以明确地看到 BIP 技术和辅助任务的独立作用以及它们的组合效应。实验结果表明,这两个模块均能显著提升模型性能,而当两者结合使用时,性能提升最为显著,尤其是在数据增强后的大规模数据集上,表现尤为突出,有效验证了每个模块对整体系统性能的贡献。

在 S2UT 模型的推理过程中,beam_size 在 CTC(连接时序分类)解码中指定了在搜索空间中可能的解决方案的数量。选择合适的 beam_size,对于平衡解码效率与结果的准确性至关重要。具体而言,较大的 beam_size 值意味着模型将在更宽广的空间内寻找可能的输出序列,从而提高了识别最优解的概率。然而,这种策略的副作用是显著提高了计算复杂度和所需的解码时间,有时甚至会引起模型过拟合,偏离预期的正确结果。为了在速度和准确性之间取得平衡,需要通过对比实验来选择合适的 beam_size。通过实验,可以评估不同 beam_size 下模型的表现,找到一个能够在合理的时间内提供最佳解的参数。通过实验(详见表 4),最终将 beam_size 设置为 10,这确保了解码过程在可接受时间内完成的同时,还能保持较高的准确率,实现了对效率和性能的双重优化。

表 4　不同 beam 对推理结果的影响

	BLEU			
	原始数据集(30h)		Tibet-AT(148h)	
	valid	test	valid	test
Beam=1	23.38	22.48	77.63	76.77
Beam=2	22.76	21.84	78.56	77.59
Beam=4	21.63	20.35	78.93	78.34
Beam=6	21.16	20.02	79.10	78.67
Beam=10	20.73	19.74	79.26	78.84

注:BLEU 分数以百分比形式(%)表示,即原始 BLEU 值 × 100%。

3　结束语

该文研究了使用自监督离散单元作为目标来训练直接 S2ST 模型。研究了在低资源多模态语言对场景下的模型训练,在开源的基线模型的基础上,替换了原有的 Hubert 模型,并使用双边扰动对替换后的 Hubert 模型进行微调,此外,为了进一步提高翻译的准确性,在模型中引入了辅助任务,该工作证明了端到端语音翻译模型在低资源远距离语言对场景下的可行性。实验表明,提出的模型与基线端到端 S2ST 模型相比,具有最好的实验效果。此外,尽管该研究聚焦于藏汉语音翻译任务,提出的方法并未依赖于藏文的特定语言学特性,而是通过数据增强与多模态特征分离技术对

小样本语言的普遍问题进行了有效建模。因此,该方法具有较强的通用性,理论上可扩展应用于其他小语种语言的语音翻译任务。在未来研究中,笔者将进一步扩充藏汉语音翻译的数据集,并验证该方法在其他小语种翻译任务中的适用性,同时探索其在跨语言多模态场景中的扩展能力,以进一步提升端到端语音翻译模型的性能。

参考文献:

[1] LAVIE A,WAIBEL A,LEVIN L,et al. JANUS-III:speech-to-speech translation in multiple languages[C]//Proc of the IEEE international conference on acoustics,speech,and signal processing (ICASSP). Munich:IEEE,1997.

[2] WAHLSTER W. Verbmobil:foundations of speech-to-speech translation[M]. [s. l.]:Springer,2000.

[3] NAKAMURA S,MARKOV K,NAKAIWA H,et al. The ATR multilingual speech-to-speech translation system[J]. IEEE Trans on Audio,Speech,and Language Processing,2006,14(5):1309-1321.

[4] JIA Y,WEISS R J,BIADSY F,et al. Direct speech-to-speech translation with a sequence-to-sequence model[J]. arXiv:1904.06037,2019.

[5] KANO T,SAKTI S,NAKAMURA S. Transformer-based direct speech-to-speech translation with transcoder[C]//Proc of the IEEE international conference on spoken language processing (SLT). Virtual Conference:IEEE,2021:958-965.

[6] DONG Qianqian, YUE Fengpeng, KO T,et al. Leveraging pseudo-labeled data to improve direct speech-to-speech translation[J]. arXiv:2205.08993,2022.

[7] LEE A,CHEN P J,WANG C,et al. Direct speech-to-speech translation with discrete units[J]. arXiv:2107.05604,2021.

[8] WEI Kun. Joint pre-training with speech and bilingual text for direct speech-to-speech translation[C]//Proceedings of IEEE international conference on acoustics,speech and signal processing (ICASSP). Rhodes:IEEE,2023:1-5.

[9] CHOROWSKI J,WEISS R J,BENGIO S,et al. Unsupervised speech representation learning using wavenet autoencoders[J]. IEEE/ACM Trans on Audio,Speech,and Language Processing,2019,27(12):2041-2053.

[10] CHUNG Yu-An,HSU Wei-Ning,TANG Hao,et al. An unsupervised autoregressive model for speech representation learning[J]. arXiv:1904.03240,2019.

[11] LEEA,GONG Hongyu,DUQUENNE P A,et al. Textless speech-to-speech translation on real data[J]. arXiv:2112.08352,2021.

[12] HUANG Rongjie,LIU Jinglin,LIU Huadai,et al. TranSpeech:speech-to-speech translation with bilateral perturbation[EB/OL]. (2023)[2024-10-29]. https://github.com/Rongjiehuang/TranSpeech.

[13] TJANDRA A,SAKTI S,NAKAMURA S,et al. Speech-to-speech translation between untranscribed unknown languages[J]. arXiv:1910.00795,2019.

[14] ZHANG Chen,TAN Xu,REN Yi,et al. Uwspeech:speech to speech translation for unwritten languages[J]. arXiv:2006.07926,2020.

[15] 张 敏. 基于 BERT 模型的汉藏跨语言转换研究[D]. 兰州:西北师范大学,2023.

[16] 王 瑞. 基于自动语种识别的汉藏双语跨语言语音转换研究[D]. 兰州:西北师范大学,2022.

[17] HSU W N,BOLTE B,TSAI Y H H,et al. HuBERT:self-supervised speech representation learning by masked prediction of hidden units[J]. arXiv:2106.07447,2021.

[18] POLYAK A,ADI Y,COPET J,et al. Speech resynthesis from discrete disentangled self-supervised representations[J]. arXiv:2104.00355,2021.

[19] CUI Chenye,REN Yi,LIU Jinglin,et al. Varietysound:timbre-controllable video to sound generation via unsupervised information disentanglement[EB/OL]. (2022)[2018-03-29]. https://ieeexplore.ieee.org/document/10096353.

[20] BAEVSKI A,AULI M,MOHAMED A. Effectiveness of self-supervised pretraining for speech recognition[J]. arXiv:1911.03912,2019.

[21] VASWANI A,SHAZEER N,PARMA N,et al. Attention is all you need[C]//Proc of advances in neural information processing systems. Long Beach:NeurIPS,2017:5998-6008.

[22] SYNNAEVE G,XU Qiantong,KAHN J,et al. End-to-end ASR:from supervised to semi-supervised learning with modern architectures[J]. arXiv:1911.08460,2019.

[23] KONG J,KIM J,BAE J. HiFi-GAN:generative adversarial networks for efficient and high fidelity speech synthesis[C]//Proc of advances in neural information processing systems. Virtual Conference:NeurIPS,2020:17060-17071.

[24] REN Yi,HU Chenxu,TAN Xu,et al. Fastspeech2:fast and high-quality end-to-end text to speech[J]. arXiv:2006.04558,2020.

[25] ZHANG Binbin,LV Hang,GUO Pengcheng,et al. Wenet-Speech:a 10000+ hours multi-domain Mandarin corpus for speech recognition[C]//Proc of the IEEE international conference on acoustics,speech,and signal processing (ICASSP). Toronto:IEEE,2021.

[26] 赵小兵,刘佳洛,周毛克,等. 藏汉语音翻译数据集[J]. 中国科学数据:中英文网络版,2024,9(4):21-29.